Education
Endowment
Foundation

Increasing Pupil Motivation

Evaluation Report and Executive Summary

October 2014

Independent evaluator:

Institute for
Fiscal Studies

Luke Sibieta

Ellen Greaves

Barbara Sianesi

The Education Endowment Foundation (EEF)

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- Identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- Evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale;
- Encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust, as lead charity in partnership with Impetus Trust (now part of Impetus-The Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

James Richardson
Senior Analyst
Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

p: 020 7802 1924
e: james.richardson@eefoundation.org.uk
w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the Institute for Fiscal Studies (IFS). Luke Sibieta, Programme Director of the Education and Skills sector at IFS, led the impact evaluation.

Contact details:

Luke Sibieta
Programme Director, Institute for Fiscal Studies

7 Ridgmount Street
London
WC1E 7AE

p: 020 7291 4800
e: luke_s@ifs.org.uk

Contents

Executive Summary 2

Introduction 7

Methodology 10

Impact evaluation 19

Process evaluation 34

Conclusion 36

References 38

Appendices 40

Executive summary

The project

‘Increasing Pupil Motivation’ was designed to improve attainment at GCSE by providing incentives to increase pupil effort in Year 11. Two schemes for incentivising pupil effort were implemented. The first provided a financial incentive, where pupils were told they had £80 at the beginning of each half-term. Money was deducted if they did not reach the threshold in four measures of effort: attendance, behaviour, classwork and homework. The second provided an incentive of a trip or event. Pupils were allocated a certain number of tickets at the start of term and lost them if they failed to meet targets on the same set of four effort thresholds. Pupils that retained enough ‘tickets’ were rewarded with an event, chosen by pupils in the year group at the start of the school term.

Pupil effort was monitored by the schools involved in the intervention, but the design and development of the incentive schemes was undertaken by the project team at the University of Bristol in the first four half-terms of the 2012/13 academic year.

The target population was relatively deprived schools, classified by schools with an average pupil IDACI (Income Deprivation Affecting Children Index) score in the highest 10% in England. 279 eligible schools were invited to participate by the project team, with 84 schools indicating an initial willingness to participate. 63 schools agreed to participate in the intervention after an initial training event in July 2012; they then formed the final set of experimental schools.

Key conclusions

1. **Event Incentives –** There is no evidence of a significant positive impact of event incentives on GCSE attainment in Maths, English or Science.
2. **Financial Incentives –** There is no evidence of a significant positive impact of financial incentives on GCSE attainment in Maths, English or Science.
3. There is a statistically significant improvement in classwork effort across English, Maths and Science for the financial incentive treatment. There is no evidence of impact on behaviour, attendance or homework effort. This may suggest that even when there is a marked improvement in effort in classwork, this does not translate into higher GCSE attainment.
4. There was a positive impact of both the event and financial incentives on GCSE Maths for pupils with low levels of prior attainment – equivalent to about one quarter of a GCSE grade in Maths, although this is not statistically significant for the financial incentive treatment.
5. Schools found it difficult to organise and pay for events before they knew how many pupils were likely to meet their targets. Schools should also consider the cost of monitoring and providing feedback about pupil effort.
6. Further research should explore the level of incentive required to induce pupil effort, and the long-term impact of such schemes. Further research might also be needed to see if there are any adverse effects if schools just decided to incentivise one subject (e.g. Maths) or just one group of pupils (e.g. those with low levels of prior attainment). Additionally, future studies should explore why incentives appear to change classwork effort but do not necessarily translate into higher attainment. The relationship between improved pupil effort and its impact on attainment should be examined in greater detail.

What impact did it have?

The estimated effects of the financial and event incentives are shown in Tables 1 and 2 for GCSE Maths and English, respectively (results for Science are shown in the main report). There is no secure evidence from this evaluation that financial incentives have an impact on overall pupil attainment in GCSE Maths or English. The small positive effect sizes detected – 0.04 and 0.02 in Maths and English respectively – are not statistically significant from zero. There is also no evidence of a statistically significant effect on GCSE Science. This evaluation can rule out medium to large effects on pupil attainment.

There is some suggestion that when pupils were offered the incentive of an event or trip, there was a small positive impact on Maths attainment at GCSE, but this is not statistically significant from zero.[1] The effect size of 0.08 is the equivalent of approximately one month's progress or one sixth of a GCSE grade. There was no evidence of a statistically significant effect on GCSE English or Science though.

There is evidence to suggest that the impact of the event incentive treatment is larger in Maths for pupils with low levels of prior attainment at Key Stage 2. The estimated impact for this incentive treatment is 0.13, significant at the 5% level, which is the equivalent of approximately two months' progress, or one quarter of a grade in GCSE Maths. For English, the estimated effects of both sets of incentive for pupils with low prior attainment were also positive, but not statistically significant. For Science, the effects were closer to zero.

A secondary outcome measure was pupil effort. For the financial incentive, there was a positive and statistically significant improvement in classwork in English, Maths and Science (at the 5% level). For the event incentive, there was also a positive impact although this was not statistically significant. Across the other measures of effort, there was no secure evidence of a positive impact. Combined with the results on attainment, this suggests that improvements in classwork is the main area where effort has improved as a result of the incentive schemes, but there is only some evidence to suggest this translates into an effect on attainment in the case of Maths. One possible explanation is that classwork effort in Maths translates into higher GCSE attainment than classwork effort in English or Science.

There was no process evaluation commissioned as part of this independent evaluation.

[1] Statistical significance is reported here at the 5% level, which is usual practice in education research. However, as the analysis in the main report shows, some effect sizes are significant at the 10% level, which means there is a one in ten chance that an effect of this size or larger could be observed in this sample of schools when in fact there is no true impact of the intervention. If a 10% level of significance is used, the findings should be greeted with a greater degree of cautious interpretation than significance at the 5% level. Further details and explanation are provided in the main report.

Table 1: Summary of estimated impacts of financial incentives

Group **GCSE Maths Points Score**	**Number of pupils (schools)**	**Effect size (95% confidence interval)**	**Estimated months' progress**	**Is this finding statistically significant?***	**Evidence strength****	**Cost of Approach*****
All pupils	7,730 (48)	0.04 (-0.06, 0.13)	+1[2]	No	🔒🔒🔒🔒🔒	£££
Pupils eligible for free school meals	2,934 (48)	0.06 (-0.06, 0.18)	+1	No	▨	▨
Low prior attainment (Key Stage 2)	2,551 (48)	0.14 (-0.02, 0.29)	+2	No	▨	▨
Group GCSE English Points Score						
All pupils	7,730 (48)	0.02 (-0.08, 0.12)	0	No	🔒🔒🔒🔒🔒	£££
Pupils eligible for free school meals	2,934 (48)	0.03 (-0.09, 0.15)	0	No	▨	▨
Low prior attainment (Key Stage 2)	2,551 (48)	0.17 (-0.04, 0.27)	+2	No	▨	▨

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). The effect of each treatment is estimated separately compared with the full set of 33 control schools, i.e. the effect of the financial incentives represents the 15 financial incentive schools compared with the 33 control schools. All 63 schools are used in this analysis. Low prior attainment is defined as achieving below 4 at Key Stage 2 in Maths, English or Science.
**For more information about evidence ratings, see Appendix D in the main evaluation report.
Evidence ratings are not provided for sub-group analyses, which will always be less secure than overall findings
***For more information about cost ratings, see Appendix E in the main evaluation report.

[2] *Since this report was published, the conversion from effect size into months of additional progress has been slightly revised. If this result was reported using the new conversion, it would be reported as 0 months of additional progress rather than +1. See* ***here*** *for more details.*

Table 2: Summary of estimated impacts of event incentives

Group **GCSE Maths Points Score**	**Number of pupils (schools)**	**Effect size (95% confidence interval)**	**Estimated months' progress**	**Is this finding statistically significant?***	**Evidence strength****	**Cost of Approach*****
All pupils	7,980 (48)	0.08 (-0.01, 0.17)	+1	No	🔒🔒🔓🔓🔓	£
Pupils eligible for free school meals	3,190 (48)	0.09 (-0.02, 0.20)	+1	No	▨	▨
Low prior attainment (Key Stage 2)	2,619 (48)	0.13* (0.02, 0.24)	+2	Yes	▨	▨
Group GCSE English Points Score						
All pupils	7,980 (48)	0.04 (-0.08, 0.16)	+1	No	🔒🔒🔓🔓🔓	£
Pupils eligible for free school meals	3,190 (48)	0.08 (-0.06, 0.21)	+1	No	▨	▨
Low prior attainment (Key Stage 2)	2,619 (48)	0.10 (-0.06, 0.25)	+2	No	▨	▨

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). The effect of each treatment is estimated separately compared with the full set of 33 control schools, i.e. the effect of the event incentives represents the 15 event incentive schools compared with the 33 control schools. All 63 schools are used in this analysis. Low prior attainment is defined as achieving below 4 at Key Stage 2 in Maths, English or Science.

**For more information about evidence ratings, see Appendix C in the main evaluation report.

Evidence ratings are not provided for sub-group analyses, which will always be less secure than overall findings

*** For more information about cost ratings, see Appendix D in the main evaluation report.

How secure is the finding?

Previous research has focused on the provision of incentives that are awarded directly for test scores. Most studies found positive effects on test scores, but these positive findings are not universal. Relatively few studies have examined the effect of rewarding pupils specifically for effort tasks, rather than test score performance. One study by Fryer (2011) examined the effect of paying pupils to read books and finds little positive effect on test scores. However, as the author acknowledged, this study had some limitations. Therefore, there is currently no clear finding in the literature surrounding the effects of direct incentives on pupil performance and very little work that examines the effect of incentives for pupil effort rather than performance.

The evaluation was designed as a cluster randomised controlled trial, with randomisation at the school level, and two treatment groups of schools and one control group of schools. In one treatment group, Year 11 pupils received financial rewards for the successful completion of effort tasks. In the second treatment group Year 11 pupils received a non-financial reward in the form of a trip or event for effort tasks.

The evaluation was run as an effectiveness trial. Effectiveness trials seek to test if an intervention will work under typical conditions across a number of schools in different settings.

The design of the trial is sound. However, there are two limitations that affect the interpretation of the findings. First, the incentives were combined with a system of feedback in both treatment groups – pupils were informed of their performance and how close they were to achieving their targets. It is therefore not possible to estimate the effects of providing incentives and providing feedback separately; we must instead estimate the combined effect of the provision of incentives and feedback on pupil effort and attainment.

Second, a number of control schools (18 out of 33) did not provide effort data to the project team, which implies that pupil effort was not monitored in these schools. This subset of control schools therefore differed from treatment schools in the provision of incentives for effort, feedback and monitoring of pupil effort. To mitigate this potential problem, a robustness check was conducted to estimate the effect of both the financial and event incentive treatment relative to the subset of control schools that monitored pupil effort (those that submitted effort data to the project team). The estimated impact of each incentive treatment is not lower when we drop controls who did not submit effort, though it is higher, suggesting that schools who dropped out may be different in unobservable ways.

It is also important to acknowledge that in this trial the cost of the incentives was met by the project budget, and not directly from school budgets, and it was teachers who recorded pupils' effort. It is therefore possible that teachers were more generous in allowing pupils to reach thresholds because they knew the money was not coming directly from school budgets. However, this is likely to be the case only in marginal decisions.

How much does it cost?

The cost of the financial incentives represented a maximum of £320 per pupil (if all targets for effort were met). The average cost per pupil was approximately £225. The cost of the event incentive treatment was the budget given to schools for organising an event at the end of each term. The budget allocated by the project team was about £80 per pupil to cover both terms. In addition to this, all schools were given an additional £2,000 to cover the expected cost of monitoring pupils' effort, although the true cost of monitoring pupils was not calculated.

One important caveat to these figures is that this intervention focused on one particular level of incentives. Schools could in principle choose to offer lower or higher levels of incentives for pupil effort and the cost would naturally change as a result. However, more evidence would be needed in order to calculate the likely effect of offering lower or higher levels of incentives.

Schools should also consider the cost in staff time and effort of organising and delivering the chosen incentive. For example, the project team reported some difficulties in paying the financial rewards in general. The project team also reported that the organisation of events at the end of term was operationally complex as deposits often had to be paid, even before schools knew how many pupils were likely to attend.

Introduction

Intervention

The aim of this intervention was to test whether the provision of financial and non-financial incentives for effort tasks improves GCSE performance. The intervention was tested on pupils in the final year of compulsory schooling (also known as Year 11, when pupils are typically aged 15-16) across 63 schools during the first two terms of the academic year. Schools were allocated to one of three groups of schools. Pupils in the control group received no incentives for pupil effort, but pupil effort was monitored in the same way as the two treatment groups. Pupils in one treatment group received financial rewards at the end of each half-term depending on their effort. Pupils in the other treatment group were able to attend an event at the end of each term if their effort met a certain threshold. Both treatments aimed to incorporate loss aversion (the idea that individuals dislike losses more than they like gains of the same value). These targets and thresholds were determined by the project team and were constant across all pupils and schools.

Background evidence

Previous studies have examined the impact of providing financial incentives to students. These studies have usually tested the effect of paying financial rewards to students depending on their test scores. Most papers have found positive effects on test scores (O'Neil et al., 1997; Jackson, 2010; Angrist and Lavy, 2009; Bettinger, 2010; Braun et al., 2011; Levitt et al., 2011), however, these positive findings are not universal (Fryer, 2011; Baumert and Demmrich, 2001). Furthermore, most of the existing evidence suggests that the way the incentives are designed and framed matters, and that they could be differentially effective across groups of students and subjects. Levitt et al. (2011) and Braun et al. (2011) find that timing matters, with immediate rewards more effective than delayed rewards, suggesting that students are impatient (or weight the present more than the future). Levitt et al. (2011) also argue that rewards framed as losses have consistently larger effects than those framed as gains (consistent with the idea of loss aversion). Angrist and Lavy (2009) find bigger effects for girls, while Levitt et al. find the reverse. Both Levitt et al. and Bettinger (2010) find larger effects for test scores in Maths than other subjects.

One might think that directly rewarding pupils for effort tasks could have positive implications for pupil attainment, particularly given recent evidence that implicitly finds that short-term effort is an important determinant of student performance in high-stakes exams (Metcalfe et al., 2011). Comparatively few studies have examined the effect of rewarding pupils specifically for effort tasks, rather than test score performance, however. The only existing paper to look at this is Fryer (2011), who examines the effect of paying pupils to read books and finds little positive effect on test scores. The Fryer (2011) study was relatively underpowered, however, and it was not possible to estimate the impact of the number of books read.

There is also a wider theoretical and empirical literature looking at the role of financial incentives in education and public services more generally (for a review see Burgess and Ratto, 2003). One recurring theme in this literature is the concern that financial incentives could crowd out intrinsic motivation (e.g. Deci and Ryan, 1985; Tirole and Benabou, 2006). Extrinsic rewards could also increase intrinsic motivation if they provide some external re-enforcement (Muralidharan and Sundararaman, 2011), i.e. in this intervention, pupils could become more motivated through knowing that their effort is valued and rewarded.

There is also an emerging literature on non-financial and status incentives, which argues that non-financial rewards may be more effective for intrinsically motivated workers (Ashraf et al., 2014; Kosfeld and Neckermann, 2011; Besley and Ghatak, 2008). Non-financial rewards could therefore be highly effective if intrinsic motivation is an important determinant of performance, and could be more cost-effective than financial incentives. Levitt et al. (2011) compare financial and non-financial rewards within the same intervention. They find that financial rewards for test scores are more effective for older students, but non-financial and financial incentives are equally effective for younger students.

There is currently little evidence on the incentives offered by schools. A brief survey of schools conducted by the project team found that 18 out of 33 schools that responded used some form of incentives at that time, with all but one offering trips or trophies as a reward (rather than financial incentives).

In summary, there is already a body of evidence finding positive effects of financial incentives for pupils' performance on tests, as well as some limited evidence to suggest that the design of the incentive scheme matters (with rewards framed as losses and those more immediate seeming to have higher impact). There is little evidence that looks specifically at rewards for effort tasks or that compares financial and non-financial rewards for such effort tasks. The comparison with non-financial incentives is important as these are already used by many schools, though with little empirical evidence to guide implementation or design decisions. This intervention was designed in order to fill these gaps in current knowledge and offer schools reliable empirical evidence on whether offering such incentives schemes is likely to increase pupil effort and attainment and be cost-effective. However, it is important to remember that the incentives schemes in this intervention were intended to be additional to what schools were already doing, and thus cannot test the effectiveness of the small schemes some schools were already offering.

Evaluation objectives

The main research questions for this evaluation are as follows:

1) Does the provision of financial or non-financial incentives for effort tasks improve pupil attainment measured at GCSE?
2) Does the provision of financial or non-financial incentives for effort tasks improve pupil effort?
3) Does the provision of financial or non-financial incentives for effort tasks improve pupil effort and attainment for disadvantaged pupils in particular?

Evaluation team

Lead researcher from IFS:
Luke Sibieta, Programme Director at IFS

Supported by:
Ellen Greaves, Senior Research Economist at IFS
Barbara Sianesi, Senior Research Economist at IFS

Project team

The project was led by Professor Simon Burgess at the University of Bristol and also consists of Rebecca Allen (Institute of Education), Steven Levitt (University of Chicago), John List (University of Chicago), Robert Metcalfe (University of Chicago) and Sally Sadoff (University of Chicago). The University of Bristol was responsible for designing the programme, recruiting schools, implementing the intervention and data collection.

The project team are also working on a detailed evaluation of the impact of incentives on pupil attainment and effort tasks. A draft of this was made available to the evaluation team following the initial draft version of this evaluation report. We therefore indicate where our reading of their draft working paper has informed our analysis.

Ethical review

Ethical approval was granted by University of Bristol's School of Economics, Finance and Management Research Ethics Committee. The Centre for Market and Public Organisation received confirmation of approval on 16 May 2012.

It was assessed and approved under the University's Research Registration Process (registration number 1778) for the purposes of insurance and indemnity; compliance with the data protection and safeguarding requirements; informed consent and the unlimited right to withdrawn from the project. Full registration of the study was confirmed on 18 July 2012. Ethical approval is an element of study registration.

Ethical approval was not sought by the evaluation team (IFS) as the evaluation protocol followed that of the project team.[3] Nevertheless, IFS researchers are required to adhere to the Economic and Social Research Council's Ethics Framework, and the Social Research Association's Ethical Guidelines. IFS researchers are also required to adhere to the IFS Information Security guidelines and the IFS Information Classification and Handling Policy, both of which comply with the international standard for data security (ISO27001)

An information session for all pupils in the schools allocated to one of the two treatment groups was given in September 2013. All pupils were given the option to opt-out of the intervention.

[3] This intervention was the first funded by EEF where an evaluation was commissioned from both the project team and the independent evaluation team, and as such the procedures for ethical review, randomisation and evaluation protocol were not fully established.

Methodology

Trial design

The trial was designed as a cluster randomised controlled trial, with randomisation at the school level, and two treatment groups of schools and one control group of schools. In one treatment group Year 11 pupils received financial rewards for the successful completion of effort tasks. In the second treatment group Year 11 pupils received a non-financial reward in the form of a trip or event (see below for the precise details of these schemes and the required effort tasks). Randomisation to each incentive scheme was at the school rather than pupil-level in order to avoid potential biases caused by spillover effects across pupils within the same school.[4]

The trial focused on pupils in Year 11 as this is the final year before the high-stakes GCSE exams taken at the end of compulsory schooling in England, and previous evidence suggests that effort during this final year could be an important determinant of exam success (Metcalfe et al, 2011).

There were two changes to the original design of the intervention. First, the incentive offered to all schools participating was increased to £2,000 from £1,000 as the costs of data collection for schools (monitoring and reporting pupil effort) were higher than originally expected. This decision was made after schools had been allocated to a particular treatment or control group. Second, the number of schools involved was increased beyond that envisaged in the original design. The original design specified 14 schools in each treatment group and control group (42 schools in total). Following a highly successful recruitment phase, this was increased to 15 treatment schools in each group and 33 control schools (63 schools in all).

Eligibility

The intervention was focused on improving attainment among disadvantaged pupils in Year 11. The group of eligible schools was therefore defined as a set of relatively deprived secondary schools across England. This was determined by the average neighbourhood deprivation of their pupils: eligible schools were the most disadvantaged ten percent.[5] All Year 11 pupils in these schools were eligible for the intervention.

These 279 schools were invited to participate by external consultants employed by the project team. Schools were approached by a combination of phone calls and letters. There were 84 schools who indicated they were willing to participate. This went down to 63 schools after the initial information and training event in July 2012,[6] who then formed the final set of intervention schools. Schools were not excluded from the intervention if they already operated some form of non-financial incentive scheme.

[4] This design aided recruitment as some schools were often pleased to learn that all pupils in the same year group would be treated as they had concerns about fairness and horizontal equity should the randomisation have occurred within school.

[5] The measure of average neighbourhood deprivation of pupils in the school was formed from the Income Deprivation Affecting Children Index (IDACI) for each pupil at the school. The only exceptions to the eligibility criteria were the exclusion of schools in special measures, schools about to close down, schools with pupils aged under 11 and the exclusion of one school with a very high profile reputation (as it already attracts a large amount of research interest).

[6] Some training was offered at this initial information event on collecting effort data. This was done owing to significant time pressure resulting from the experiment's planned start date in September 2012.

Consent to be involved in the intervention was sought at the school level (i.e. being willing to participate in the intervention) and at the pupil level (pupils were given the option to opt out at any stage). See the previous section on 'Ethical review' for further details.

Intervention

The main purpose of the intervention was to test the effectiveness of offering different types of incentives for the completion of effort tasks on pupil attainment at GCSE level. However, apart from the incentives offered, there were also a number of other differences across treatment and control groups that should be noted when interpreting the impact estimates. These differences are summarised in Table 2 and discussed in turn below.

Table 3: Aspects of treatment and control groups

Aspects of treatment	T1	T2	C
Existing reward schemes	At least 27%	At least 27%	At least 30%
Effort measurement	Yes	Yes	Yes (for 15 out of 33 schools)
Incentive scheme	Financial (£80 half term; £10 attendance, £10 behaviour, £30 classwork, £30 homework) and perhaps existing provision	Non-financial (event) every term requires 12 tickets out of a maximum of 16 tickets and perhaps existing provision	None or existing provision
Incorporation of loss aversion	Framed to pupils as money was 'theirs to lose'	Framed that pupils are given 8 tickets at start of each half-term, which were 'theirs to lose'	Don't know / NA
Length of time before reward	Half-termly	Termly	NA
Feedback letter to pupils	Yes	Yes	No

Before the start of the intervention, a number of schools already offered **existing incentives to pupils.** The project team conducted a small-scale survey and found that 18 out of 33 schools who responded to the survey offered some kind of incentive (mostly trophies or trips). Across the treatment and control groups, this means we know that at least 27% of each treatment group offered some kind of incentive scheme and at least 30% of control schools did so. Schools across all groups were told to continue with their existing incentive schemes over the course of the intervention. Data on whether schools offered existing incentives was also used in the randomisation procedure (see later in this section) and there is therefore no difference across treatment or control groups in terms of whether schools offered existing incentives. The only potential concern is if control schools offered new incentives in light of the intervention or if treatment schools scaled down existing schemes in light of the new ones. Schools were explicitly told not to do this as part of the initial training and information event in July 2012, but the extent to which schools followed this guidance isn't known.

The next key aspect is the **measurement of effort**, which is necessary in order to provide incentives for pupil effort. The rewards were based on effort tasks across English, Maths and Science as all pupils take these subjects in Year 11. This design ruled out any difficulties caused by differences in subject and course choice within or across schools (although pupils are able to choose different course options for Science in many schools). Effort was rewarded by meeting targets for effort across four key domains: attendance, behaviour, classwork and homework. These targets were designed to be tough, but achievable.[7]

[7] The precise definitions of the targets in each of these key domains are included in Appendix B.

All schools across treatment and control groups were expected to measure these effort tasks to ensure that the potential impact of measurement on pupil effort and attainment didn’t contribute to the estimated impact of the two incentive treatments. 18 control schools failed to provide effort monitoring data, however, suggesting that some control schools did not monitor pupils as expected. This creates some concern that differences in pupil outcomes between treatment and control schools could reflect additional monitoring in treatment schools. To address this concern in part, we perform a robustness check where we estimate the impact of the two incentive treatments using only schools that submitted effort data in the control group. The estimated effects, reported in the ‘Robustness checks’ section, are slightly more positive when we exclude schools that failed to submit effort data. This suggests that control schools who did submit effort data actually had *lower* outcomes than those who failed to submit effort data. This could be the case if monitoring had a negative impact on outcomes, but could also have resulted from schools who failed to submit data being different in unobservable ways. Our main analysis therefore still focuses on the full set of 63 schools.

The main difference across treatment and control groups was the provision of different **types of incentives,** which differed in terms of their nature, but also in terms of their incorporation of loss aversion and the timing of rewards.

Each pupil in the **financial incentive** treatment group started with an allocation of £320, and £80 of this was available each of the four half-terms. The rule for determining the financial rewards was that if pupils missed their threshold tasks (specified above), they lost rewards as follows: £10 for missing the attendance threshold; £10 for behaviour; £30 for classwork; £30 for homework. Therefore, if they met all four thresholds in the half-term, they would receive £80. Rewards were weighted towards classwork and homework by the project team due to their perceived importance in determining pupil attainment.

The non-financial incentive was the opportunity to attend an event at the end of the term. In each school, pupils collectively decided on the event and were then able to attend the event if they retained enough ‘tickets’ through the term. These events included trips to theme parks, ice skating and the Houses of Parliament. Each pupil in the **event incentive** treatment group started with an allocation of 8 ‘tickets’ per half-term. Pupils were able to attend the event at the end of each half-term if they retained enough tickets. Tickets were lost in the same ratio as financial rewards in the financial incentive treatment group: 1 for missing the attendance target, 1 for behaviour; 3 for classwork and 3 for homework. Therefore, by the end of the first term before Christmas they could have a maximum of 16 tickets. To go on the trip they needed 12 tickets. The two rounds were separate and distinct. This event reward is a threshold measure (a pupil either wins or loses). In comparison, the financial incentive was a continuous measure, so a pupil still has an incentive to exert effort even if the equivalent threshold was missed.

Both sets of rewards were designed to incorporate a degree of **loss aversion** – the idea from behavioural economics and psychology that individuals dislike losses more than they like gains. It was decided on the basis of emerging evidence in Levitt et al. (2012) suggesting that rewards incorporating loss aversion could be more effective. Loss aversion was incorporated by framing the incentives as ‘theirs to lose’ at the start of each period. Note that the design of this intervention does not allow us to directly test whether rewards incorporating loss aversion are indeed more effective.

The **length of time** between pupil effort and event reward was longer than between pupil effort and the financial reward, which may be important if pupils have a high preference for rewards in the near future. This difference was necessary as organising the events was relatively complex and organising one each half-term may have placed too much of a burden on schools.

The last key difference between treatment and control schools was the **provision of feedback letters** to pupils in treatment schools. These detailed their performance in each half-term and how this performance had determined their rewards. As a result, it is impossible to separate the effects of the provision of rewards from the provision of feedback on pupil effort and attainment: the estimated treatment effect is a combination of both.

In summary, the key difference between treatment and control groups is the provision of incentives for pupil effort across four domains and bespoke feedback letters documenting pupil effort levels. There are other differences, however, which influence the interpretation of the estimated treatment effect: it is likely that some control schools didn’t monitor pupil effort which means that the estimated treatment effect also includes this to some extent; schools may have changed their existing provision of incentive schemes after their allocation to treatment or control group.

Outcomes

The primary outcomes for this evaluation are GCSE performance in Maths, English and Science by individual pupils. Secondary outcomes include overall GCSE performance (i.e. including other subjects) and the successful completion of effort tasks.

Attainment

The primary outcome of interest is GCSE performance by individual pupils measured by performance in the subjects targeted by the intervention (Maths, English and Science). Secondary attainment outcomes of interest are measures of overall GCSE performance and measures of student effort. If successful, the intervention would clearly be expected to improve GCSE performance in the individual subjects, but may also have effects on other subjects (particularly if overall attendance improves or there are spillovers from improved behaviour across other subjects). Alternatively, students may divert effort away from subjects that are not incentivised. GCSE performance has been shown to have a high impact on later life earnings and employment (e.g. Blundell et al., 2005). The specific measures we will use are as follows:

- GCSE Maths points score
- GCSE English points score
- GCSE Science points score[8]
- Capped average point scores across eight best GCSEs (or equivalent)
- Whether children achieved a grade C or above in Maths
- Whether children achieved a grade C or above in English
- Whether children achieved the EBacc Science component (effectively a grade C in Science)
- Whether children achieved 5 or more GCSEs or their equivalent at A*-C (including English and Maths)

These measures of GCSE performance are recorded in the National Pupil Database maintained by the Department for Education, and are measured at the individual pupil level. The continuous points scores are derived as per guidelines set by the Department for Education and are then standardised relative to the national population to have a mean of zero and standard deviation of one. Binary outcomes are coded as zero if the attainment threshold is not achieved and one if the threshold is achieved.

The continuous points scores will be our main primary outcome of interest as they are the finest possible measure of GCSE performance. We will also present results using the threshold measures of attainment defined above as these represent important benchmarks for being able to access post-16 and higher education, and for employment opportunities.

GCSEs are marked independently by examiners (who do not teach at the same school as the pupil and are not aware of the intervention) and names are not visible to the marker. Some GCSEs include coursework components, which are marked within schools according to an externally set mark scheme.

[8] Calculating a points score for Science is relatively complex as pupils can take different types and qualifications. We have chosen to focus on the maximum point scored in any individual GCSE or equivalent course. Here, we gratefully acknowledge input from the project team who considered the merits of various different types of point score measures for Science. .

Effort

Our secondary outcomes are student effort levels as measured and monitored during the experiment. In particular, whether they achieved their targets on attendance, behaviour, classwork and homework as defined above:

- Whether pupils met the threshold measure of effort in attendance
- Whether pupils met the threshold measure of effort in behaviour
- Whether pupils met the threshold measure of effort in classwork
- Whether pupils met the threshold measure of effort in homework

We will analyse outcomes split by subject as well in overall terms. These outcomes are of clear interest as they represent the specific targets of the interventions and individuals will be rewarded according to their performance across these effort tasks. These effort tasks were measured and monitored by schools and teachers during the intervention.

We impose a common sample for the analysis of impact on attainment outcomes, where all outcomes of interest must be observable.[9] This ensures that differences in the estimated impact of the incentive schemes across outcomes are not due to changes in the pupils included in the sample, although some bias may be introduced. The common sample imposed for the behaviour outcomes is slightly different, as only a subset of control schools (15 out of 33) continued to monitor and provide effort data for their pupils to the project team.

Sample size

The target sample size was determined with consideration to both the budget (given the relatively high expected cost of the incentive schemes) and the minimum detectable effect size. The original sample size was set at 14 schools per treatment and control group (or about 2,400 pupils assuming 170 pupils per year group).

The first row of Table 4 below uses power calculations to determine the minimum detectable effect size under these sample sizes and how this would vary with the size of the intra-cluster correlation (ICC). With an ICC of zero the minimum detectable effect size would be 0.063, which then increases as the ICC increases, reflecting the fact that information from each additional pupil within a school provides less information relative to an additional pupil in another school. The minimum detectable effect size would be about 0.266 under an ICC of 0.1 (which is a reasonable assumption based on previous EEF interventions).

Table 4: Minimum detectable effect sizes under different assumptions and sample sizes

	Schools per treatment group	**Control schools**	**Intra-class correlation**					
			0***	0.05	0.1	0.15	0.2	0.25
Detectable effect size (original sample)	14	14	0.063	0.193	0.266	0.323	0.371	0.414
Detectable effect size (expanded sample)	15	33	0.052	0.159	0.219	0.266	0.306	0.341

Note: Assumes that the unexplained variance in the outcome variable is 0.6 after accounting for pupil characteristics and prior attainment at Key Stage 2 (the pre-test outcome in this intervention). The required significance level is 5% and power is 80% with a two-tailed test. Number of pupils per year group assumed as 170.

[9] One or more outcomes are missing for 336 cases out of a total sample of 10,649 pupils.

Following a successful round of recruitment of schools, the sample size of schools was subsequently increased to 15 schools per treatment group and 33 schools in the control group. The actual cluster sizes were also very close to the original assumption of 170 pupils (see Table 7, cohort size). The greater number of additions to the control schools was decided largely on budgetary grounds (i.e. it was infeasible to further increase the number of treatment schools). Under these new sample sizes, the minimum detectable effect size is now 0.219 under an ICC of 0.1.

There is only one other study that has directly rewarded effort tasks (Fryer, 2011), which finds effect sizes of 0.012 for reading and 0.079 for maths (both of which are statistically insignificant). The present study would not be able to be able to detect such small effect sizes. Furthermore, Fryer (2011) describes his study as underpowered as it was only able to detect effect sizes over 0.15. Table 4 suggests that the intervention in the present study is of lower power than this.

As shown in Appendix Table A1, we estimate that the ICC for Maths, English and overall GCSE points turned out to be very close to 0.1 without controls for pupil and school characteristics, and was less than 0.05 conditional on pupil and school characteristics. This means that our actual minimum detectable effect size was less than 0.16 for these outcomes. The ICC for Science was higher, however, perhaps reflecting the influence of schools on Science course choices (about 0.19 without controls and 0.14 with controls), leading to a much higher minimum detectable effect size.

Randomisation

The randomisation was performed by the project team across the final group of 63 secondary schools that were willing to participate. This was the first EEF project led by an academic project team and so the division of responsibility between the project and evaluation team was not fully established: randomisation is now performed by the independent evaluator. Randomisation was completed in August 2012, after the recruitment phase, which should help minimise any selection bias, but before pupils consented to be in the intervention. The broad process the project team chose for the randomisation was as follows (each step is discussed in more detail below):

1. Form sampling blocks (or strata)
2. Form triplets of schools within sampling blocks
3. For 15 randomly chosen triplets, randomly assign each of the three schools within the triplet to each of the treatment and control groups
4. Assign the remaining schools to the control group

Step 1

The 63 schools were divided into 6 **sampling blocks** determined by whether schools currently had some form of incentive scheme and the largest ethnic group in the school (White, Asian or Black). These blocks were chosen on the basis that these would have the largest impact on pupil attainment (given the large impact of ethnicity on pupil attainment, Wilson et al., 2011) and the potential impact of the treatment (schools with existing incentives schemes might experience different impacts to those without, in particular if they are better able to co-ordinate the scheme adopted in the intervention, or the intervention scheme displaces the existing scheme).

Table 5: Number of schools by block.

Blocking Variables		Schools
Largest broad ethnic group	**Own reward scheme?**	
White	No or missing	24
Asian	No or missing	11

Blocking Variables		Schools
Black	No or missing	10
White	Yes	10
Asian	Yes	5
Black	Yes	3

Step 2

Within each sampling block, triplets of schools were formed. This was chosen on the grounds of evidence showing that pairwise sampling made estimates more efficient (Bruhn and McKenzie, 2009; Wooldridge, 2009). It should also minimise the potential impact of attrition bias as triplets uncompromised by attrition could be analysed separately to provide an unbiased estimate with lower power. However, only 19 triplets could be formed using this process as not all of the sampling blocks were divisible by 3, leaving 6 schools remaining. These 6 schools were randomly determined within each sampling block prior to the formation of triplets.

Step 3

15 triplets from the original 19 triplets were randomly selected. Within each of these 15 triplets, the three schools were randomly allocated to one of two treatment groups and the control group.

Step 4

The remaining six schools not assigned to any triplet were allocated to the control group. The 3 randomly selected triplets (9 schools) were also allocated to the control group.

The characteristics of treatment and control schools were then compared by the project team to assess the balance of groups. If the lowest p-value across treatment and control groups (in terms of any school characteristic) was less than 10% then the assignment was disregarded. A completely new random assignment was drawn until there were no p-values less than 10%. As a result, there were no significant differences in terms of school or intake characteristics at the point of randomization..

What are the implications of this process of randomization for the comparability of treatment and control groups? First, schools that were allocated to one of the 15 triplets are randomized into each of the treatment groups and control group. We would expect the balance of school level covariates across these schools to be good, given the process of stratification. Second, schools that were allocated as one of the 6 remainder schools (step 2) or to one of the four triplets (step 3) were randomly determined. However, these schools could be systematically different from the other triplets: by chance this small number of schools could have had systematically higher or lower prior GCSE attainment, for example. Our main analysis thus focuses on the full set of 63 schools, but we also perform robustness checks by excluding the 6 remainder schools and the four extra triplets. The results are largely unchanged. Third, re-randomization (the fact that the random assignment was re-run until balance was achieved) has implications for the analysis. Unfortunately, the implications of this for analysis and inference are not yet well understood and the most up-to-date guidance (Bruhn and Mckenzie, 2009; Scott et al., 2002) suggests that the most practical approach is to control for all covariates used in the randomization.

Analysis

Attainment (primary outcome)

Our main methodology for estimating the effect of the intervention on the primary outcomes (GCSE performance in Maths, English and Science) is to compare outcomes of treatment and control groups at the pupil level after accounting for the pre-test (Key Stage 2 attainment) and pupil and school characteristics.[10] All analysis is undertaken on an intention to treat (ITT) basis. Given that randomization was at the school level, we cluster standard errors at the school level. To account for the sampling blocks, we also include the sampling blocks as control variables. This conforms with EEF guidance for interventions of this type.

The **main method** we use to account for the pre-test and pupil and school characteristics is fully-interacted linear matching (FILM). FILM allows the effect of the treatment to vary linearly with the pre-test and pupil and school characteristics, which means that it is more flexible than ordinary least squares (OLS) regression. The estimates are also more precise than both OLS and propensity score matching. The additional flexibility and greater precision are our main reasons for using FILM as our preferred methodology.

Simple comparison of pupil attainment between treatment and control groups (i.e. not accounting for the pre-test or pupil characteristics), OLS, FILM and propensity score matching should all in principle provide unbiased estimates of the effect of the intervention if the randomization has been successful. We therefore perform **robustness checks** by comparing our estimates of the treatment effects across all four methodologies. The results are, reassuringly, largely unchanged.

As a further robustness check, we also perform analysis at the **school level** and compare these with our pupil level estimates. This is undertaken for raw comparisons and using OLS. Outcomes at the school level are generated by calculating the average within each school. For this analysis we weight by the size of the year group using analytical weights. These school-level results give a similar pattern to the pupil-level results, though (as one would expect) they are slightly less precise.

To account for the **randomization process**, we include indicators for the sampling blocks in our main specification. As robustness checks we: (1) restrict the sample to the 57 schools that were allocated to a triplet; (2) restrict the sample to the 45 schools that were part of the 15 triplets that were randomized into treatment and control groups and include indicators for each of the 15 triplets; (3) restrict the sample to the schools that submitted effort data.

As all schools can be followed using the National Pupil Database, there was never any potential for **dropout** to affect our ability to compare the attainment of all pupils in all treatment and control schools. We can thus compare pupils in the 15 schools in each treatment group to the full set of 33 control schools (and each chosen subset of control schools in our robustness checks). There was also no dropout of treatment schools from the intervention, meaning that none of the original triplets is compromised in this sense. Some control schools implicitly dropped out by not submitting their effort task data. This dropout will bias our results if schools that dropped out did not monitor their pupils' effort (which seems likely) and monitoring affects pupil performance. However, comparisons of outcomes and characteristics across control schools who submitted data and those who dropped out suggest few differences.

We also undertake some limited **sub-group analysis**. In particular, we examine the impact on attainment among pupils eligible and registered for free school meals (FSM) and pupils with low prior

[10] In particular, we control for gender, FSM eligibility, whether pupils have EAL, whether pupils have a statement of SEN, whether pupils have SEN with no statement, ethnic group (minor categories included in the National Pupil Database), quarter of birth and fine points score in KS2 English, Maths and Science. In addition, we control for school-level variables for prior measures of value added across English, Maths and Science, capped GCSE points score and whether schools are in London or not. We selected these school characteristics as these are the most likely to determine current attainment and school quality. We gratefully acknowledge input from the project team in the determination of the ideal set of prior school quality variables.

attainment (those achieving Level 3 or below in one or more subjects at Key Stage 2). We again use FILM to undertake this sub-group analysis as the estimates are more precise and allow for a direct comparison with our estimates of the impact on attainment for the whole population of intervention schools.

All analysis was conducted using Stata 13. Our syntax is clearly documented and available to access from the UK data archive.

Analysis of pupil effort (secondary outcome)

The analysis of pupil effort (our secondary outcome) follows exactly the same structure as that for attainment. We focus on our preferred specification (FILM – including the same set of pupil and school covariates as in the pupil-level analysis of pupil attainment) for brevity for these secondary outcomes. The results are presented in graphical form in the main discussion which clearly shows the changes in pupil effort in each domain across the four half-terms that were monitored. The relevant appendix tables present the estimates for each treatment group in each half-term, and the comparison of effort across English, Maths and Science.

We consider the impact of the incentive schemes on pupil effort for two sub-groups of pupils: pupils eligible and registered for FSM and pupils classified as having a low level of prior attainment. This is defined as achieving below the expected level of attainment in English, Maths or Science at the end of primary school.

Process evaluation methodology

There was no process evaluation commissioned as part of this independent evaluation. Instead, the project team conducted their own semi-structured interviews with staff members at the school involved in the intervention to ascertain initial provision of incentives before the experiment, implementation issues (collecting data and paying the incentives) and schools' perception of the impact (including whether they planned to continue with the incentives after the experiment). The main findings from these semi-structured interviews are summarised very briefly in the 'Process evaluation' section, following the guidance of the project team

.

Impact evaluation

Timeline

This intervention relates to pupils in Year 11 during the academic year of 2012/13. The overall timeline for this project is shown below. The recruitment and training phase lasted from April 2012 to September 2012, with recruitment largely taking place during May and June, followed by an initial information and training event in July 2012 for interested schools. The training at this event focused on data collection and was necessary given the short timeframe before the planned start of the intervention in September.

Figure 1: Project timeline

2012	
Recruitment and Training Phase	
April–June	Recruitment of schools by project team and consultants
July	Initial training event
August	Randomization by project team
September	Follow-up training event for all schools Parents and pupils introduced to the project
Intervention Phase	
September–October	First half-term
October	Effort data collection 1 by project team Financial rewards pupils receive feedback letters and rewards Event incentive treatment pupils receive feedback letters and points won to date for Christmas event
November–December	Second half-term
December	Effort data collection 2 by project team Financial rewards pupils receive feedback letters and rewards Event incentive treatment pupils receive feedback letters and receive event reward if achieved sufficient points
2013	
January–February	Third half-term
February	Effort data collection 3 by project team Financial rewards pupils receive feedback letters and rewards Event incentive treatment pupils receive feedback letters and points won to date for Easter event
February–March	Fourth half-term
April	Effort data collection 4 by project team Financial rewards pupils receive feedback letters and rewards Event incentive treatment pupils receive feedback letters and receive event reward if achieved sufficient points
Evaluation Phase	
May–June	Pupils take GCSE and equivalent exams
December	Project team receive initial GCSE data
2014	
May	Evaluation team receive final data from project team
July	Initial draft report provided to EEF
October	Expected publication of EEF evaluation report

Randomization took place in August 2012. There was a final training event for schools in September 2012, and the intervention phase then lasted from September 2012 to April 2013. At the end of each half-term, schools submitted effort data to the project team. The project team then used this data to create feedback letters, which were sent to pupils by the start of the next half-term. Pupils in the financial incentive treatment group also received their financial reward by the start of the next half-term. Pupils in the event incentive treatment group attended the agreed event if the target was met on the date scheduled by the school (either at the end of each term or start of the following term).

Pupils took their GCSE exams in summer 2013 (Maths, English and Science GCSEs representing the primary outcome for this evaluation). The project team received initial GCSE results data in late 2013 and undertook some initial analysis. Final cleaned data (and syntax used to clean the data) was sent to the evaluation team in May 2014. This consisted of the following data merged together: data on school characteristics used during the randomization; whether schools reported having an incentive scheme in the bespoke survey by the project team; NPD data on pupils' characteristics and attainment at GCSE; the effort data submitted by schools during the intervention.

Between May 2014 and July 2014 the evaluation team created definitions of pupil and school characteristics based on raw data from the NPD and Edubase included in the dataset received by the project team, conducted the analysis and wrote the draft report. Data from the project team that was not adjusted in some way by the evaluation team includes: treatment status, sample blocks and effort data (where we rely on the data submitted to the project team by schools and cleaned by the project team).

A draft report was delivered in July 2014, with an expected publication date of October 2014.

Participants

Schools were recruited by the project team between April and June 2012. Figure 2 provides a flow chart showing how the number of schools in the intervention was determined.

The project team initially identified a group of 279 secondary schools eligible for the intervention (those in the most deprived 10% according to the average neighbourhood deprivation of all pupils in the school). This list was then provided to a group of external consultants with significant experience in contacting schools. The consultants then contacted these schools via a combination of letters and phone calls that followed an initial script designed by the project team. Schools were told that they would receive £1,000 if they participated in the intervention, which was designed to cover the costs of collecting effort data (later increased to £2,000 as the burden of collecting the data for schools was higher than expected).

Recruitment stopped once the number of willing schools reached the quota of 84 schools. These schools were then invited to an initial information and data collection training event in July 2012. After this event, a number of schools dropped out (mainly citing the potential burden of the data collection), leaving a final group of 63 schools who formed the sample of schools used in the final intervention. All Year 11 pupils in these schools were eligible for the incentives if they were in a treatment school and did not opt out of the incentives. In principle, it is possible that Year 11 pupils could have moved schools over the course of the 2012/13 academic year (the treatment phase), but this is very rare and not possible to track within the NPD data we have access to.

To illustrate the experimental context of these schools, Table 6 compares the characteristics of these 63 intervention schools with all secondary schools in England (including both maintained schools and academies). Consistent with the eligibility criteria for the intervention, schools in the intervention are clearly much more deprived than other secondary schools, with a greater share of pupils eligible for FSM. Partly because the intervention was focused on deprived pupils, schools in the intervention also have a higher share of pupils from ethnic minority backgrounds, are more likely to have English as an Additional Language and are more likely to be located in London. There is also a slightly larger share of single sex schools in the intervention.

Of these 63 schools, 15 schools were randomised into each of the financial incentives and event incentive treatment group, leaving 33 control schools. Only 15 schools were included in each treatment group owing to budget constraints. No treatment schools dropped out during the intervention. There were 18 control schools who failed to provide effort data during the intervention, effectively dropping out of the intervention. A number of these schools indicated that they had chosen to drop out because they had not been randomized into either of the treatment groups. We are still able to include these schools in our intention-to-treat analysis as GCSE performance of pupils in these schools is recorded in the National Pupil Database. As the monitoring and reporting of pupil effort may influence attainment, however, we report details of a robustness check using only schools that submitted effort data in the control group. Our analysis of pupil effort is also clearly limited to this sub-set of schools.

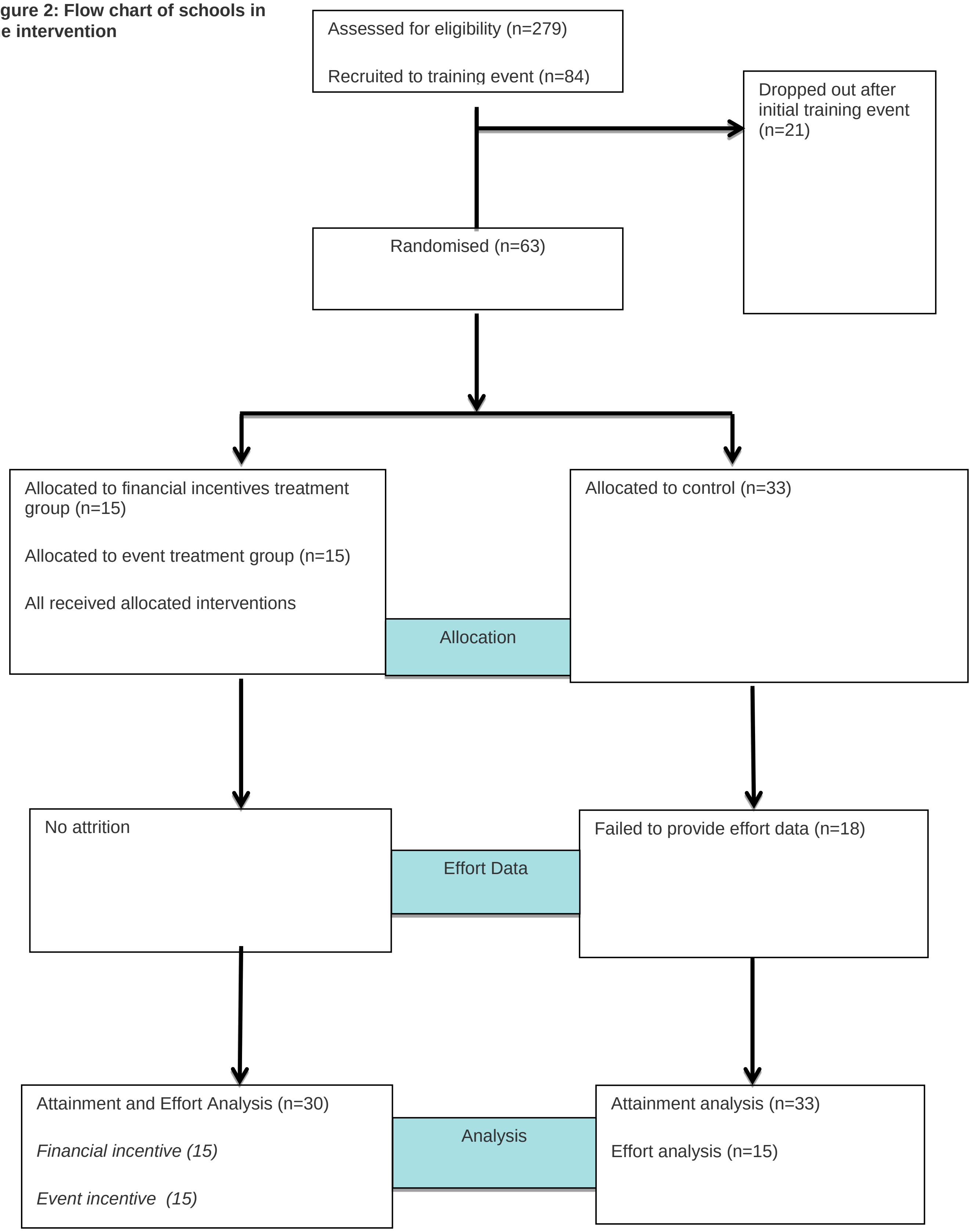
Table 6: Comparison of intervention schools with all secondary schools

Characteristic	Intervention schools	All secondary schools	Difference
Number of schools	63	3,089	
Student demographics			
Percentage who are female	0.509	0.494	0.015**
Percentage who are non-white	0.662	0.254	0.408**
Percentage of Asian ethnicity	0.221	0.088	0.133**
Percentage of Black ethnicity	0.213	0.050	0.164**
Percentage eligible for FSM	0.410	0.164	0.247**
Percentage with EAL	0.449	0.137	0.312**
Structural characteristics			
London	0.508	0.140	0.368**
Single sex school	0.159	0.124	0.035**
Academy	0.429	0.512	-0.083**
School size	924	1006	-82.3**

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Student demographics refer to the year group rather than the whole school.

In order to assess whether the dropout of control schools was random, we compare the outcomes and characteristics of control schools who stayed and dropped out of the intervention in Appendix Tables A2 and A3, respectively. This shows that the primary post-test outcomes are similar across both groups, but slightly higher among schools who continued to monitor effort (though the differences are not statistically significant). Characteristics of pupils and schools are also generally similar. The only exceptions are that schools who continued to monitor effort were less likely to be in London, more likely to have an existing incentive scheme and contain a greater share of pupils of an Asian ethnicity, though only the latter difference is statistically significant. An overall test of all the differences (a chi-squared test) suggests we should reject a null hypothesis that the financial incentive and control groups are balanced across all characteristics. However, as already stated, the differences are generally small in magnitude.

Figure 2: Flow chart of schools in the intervention


Assessed for eligibility (n=279)
Recruited to training event (n=84)
Dropped out after initial training event (n=21)
Randomised (n=63)
Allocated to financial incentives treatment group (n=15)
Allocated to event treatment group (n=15)
All received allocated interventions
Allocation
Allocated to control (n=33)
No attrition
Effort Data
Failed to provide effort data (n=18)
Attainment and Effort Analysis (n=30)
Financial incentive (15)
Event incentive (15)
Analysis
Attainment analysis (n=33)
Effort analysis (n=15)

Pupil and school characteristics

Table 7 shows how the pre-test characteristics of treatment and control schools differ, and whether these differences are statistically significant. The characteristics relate to the cohort of pupils and schools during the intervention phase of the intervention (2012/13).

Table 7 shows that there are no significant differences in the characteristics of schools across treatment and control groups in terms of their pupil composition or structural characteristics. Although not statistically significant, a number of the differences are noteworthy given their magnitude: a greater share of treatment schools are in London and there are more single sex schools in the treatment groups. There are no significant differences across treatment and control groups in prior school quality, though treatment schools generally seem to have added less value to pupils' progress across English, Maths and Science in the previous academic year (measured from KS2 to KS4).

Table 8 shows the characteristics of pupils across treatment and control schools (together with the number of pupils and schools across each group). Differences between the two treatments and the group of control schools are also shown, together with an indication as to whether these differences are statistically significant. At this stage, we calculate robust standard errors, but do not cluster at the school level. This is because we do not want to down-weight individual schools that could be driving differences across the groups (e.g. we do not want to discount differences in pupil characteristics even if it is driven by one unusual school).

There are a number of statistically significant differences across the treatment and control groups. In particular:

- There are more **male** students in the treatment groups, particularly the event incentive treatment, probably because there are more single sex schools in the treatment groups.
- The proportion of **pupils eligible for FSM** is higher in the event incentive treatment group.
- There are some clear differences in the proportion of pupils from **ethnic minority groups** (particularly in terms of pupils from different Asian backgrounds): the financial incentive treatment group contains much higher numbers of Bangladeshi pupils than the control group, but fewer pupils from Indian and Pakistani backgrounds. The event incentive treatment group is more similar to the control group, but still contains fewer pupils from Pakistani backgrounds and more from Indian backgrounds.
- Pupils in both treatment groups have higher levels of prior attainment at Key Stage 2 in both Maths and Science.

Overall, these individual differences suggest that the two treatment groups are imbalanced in terms of pupil characteristics as compared with control schools. As some differences are to be expected given the range of characteristics we consider, we also perform a chi-squared test which tests a hypothesis that the groups are balanced across all characteristics considered together. This test (with a p-value of less than 0.01) suggests that we should reject this hypothesis and that the groups are imbalanced. This should not be surprising given the relatively low numbers of schools involved. However, almost all of the significant differences become statistically insignificant when we cluster standard errors at the school level, with the exception of the number of pupils from a mixed ethnicity, where the difference remains statistically significant. Furthermore, the key issue is whether accounting for pupil and school characteristics sizeably changes our impact estimates. As we shall see in the next section, accounting for observed pre-test differences in pupil and school characteristics does not greatly change our estimates for the impact of financial incentives. This suggests that the small imbalances in pre-test characteristics we observe do not substantively affect our estimates of the impact of the financial incentives treatment.

One important caveat comes from our robustness check that uses kernel propensity score matching to estimate the effect of the two incentive treatments. In the case of the event incentive treatment, the matching estimators find it difficult to balance the pre-test observable characteristics of the event incentive treatment and control group. There are some sizeable differences in the characteristics of schools in the event incentive treatment and control group. Given that there are only a small number of schools involved, it becomes also impossible to re-weight the control group to match the treatment group along school characteristics. This creates two concerns. First, the matching estimators are likely to be biased estimates of the treatment effect. Second, and much more importantly, the scale of the differences creates some concern as to whether the event incentive treatment and control group are truly comparable. In principle, a regression-based estimate can control for observable differences in pre-test characteristics. However, it can only do so on a linear basis and this might not be sufficient when the scale of the differences is large.

Table 7: Comparison of school characteristics (analysis at school level)

Characteristic	Control group (C)	Financial incentives (T1)	Group incentives (T2)	Diff T1-C	Diff T2-C
Student demographics					
Percentage who are female	0.550	0.487	0.455	-0.063	-0.096
Percentage who are non-white	0.635	0.669	0.662	0.034	0.028
Percentage of black ethnicity	0.210	0.249	0.218	0.039	0.009
Percentage of Asian ethnicity	0.190	0.200	0.171	0.01	-0.019
Percentage eligible for FSM	0.391	0.358	0.429	-0.033	0.038
Percentage with SEN statement	0.022	0.018	0.022	-0.003	0.001
Percentage with EAL	0.444	0.445	0.480	0.001	0.036
Structural characteristics					
London	0.424	0.533	0.667	0.109	0.242
Single sex school	0.152	0.200	0.200	0.048	0.048
Academy	0.303	0.400	0.333	0.097	0.030
Has a sixth form	0.636	0.800	0.533	0.164	-0.103
School size	915.121	842.214	1004.667	-72.907	89.545
Cohort size	168.273	160.467	179.267	-7.806	10.994
Existing incentive scheme	0.303	0.267	0.267	-0.036	-0.036
Prior performance					
2012 English Baccalaureate Maths Value Added measure	1000.406	999.880	1000.487	-0.526	0.081
2012 English Baccalaureate English Value Added measure	1000.755	1000.040	999.727	-0.715	-1.028
2012 English Baccalaureate Science Value Added measure	999.848	999.260	999.387	-0.588	-0.462
2012 Total average (capped) point score per pupil	330.006	321.327	321.213	-8.679	-8.793
		Chi-squared test		0.147	0.075
		Median percentage bias		19.242	13.578

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Student demographics refer to the year group rather than the whole school.

Table 8: Comparison of pupil characteristics

	Control group (C)	Financial incentives (T1)	Event incentive treatment (T2)	Diff T1-C	Diff T2-C
Number of pupils	**5553**	**2407**	**2689**		
Number of schools	**33**	**15**	**15**		
Pupil characteristics					
Female	0.561	0.493	0.469	-0.067**	-0.092**
Eligible for FSM	0.387	0.374	0.441	-0.013	0.054**
EAL	0.475	0.462	0.491	-0.013	0.016
Statement of SEN	0.021	0.017	0.021	-0.004	0
School Action/Plus	0.237	0.224	0.272	-0.013	0.035**
White British	0.336	0.322	0.326	-0.013	-0.010
White Other	0.086	0.075	0.104	-0.011	0.018**
Black Caribbean	0.059	0.056	0.061	-0.002	0.002
Black African	0.136	0.150	0.142	0.013	0.006
Black Other	0.020	0.027	0.020	0.007	0
Asian Indian	0.048	0.017	0.060	-0.031**	0.012*
Asian Pakistani	0.094	0.066	0.046	-0.028**	-0.048**
Asian Bangladeshi	0.075	0.154	0.069	0.079**	-0.006
Asian Chinese	0.004	0.005	0.009	0	0.005**
Mixed Ethnicity	0.040	0.029	0.042	-0.011*	0.002
Other Ethnicity	0.102	0.100	0.120	-0.002	0.018*
Born in Autumn	0.253	0.249	0.268	-0.004	0.015
Born in Winter	0.245	0.257	0.247	0.012	0.002
Born in Spring	0.255	0.240	0.238	-0.015	-0.016
Born in Summer	0.248	0.254	0.247	0.007	-0.001
KS2 Eng fine point score	4.287	4.276	4.290	-0.010	0.004
KS2 Mat fine point score	4.282	4.346	4.328	0.064**	0.047*
KS2 Sci fine point score	4.553	4.587	4.588	0.035	0.035
Any missing KS2 results	0.112	0.130	0.134	0.018*	0.022**

	Control group (C)	Financial incentives (T1)	Event incentive treatment (T2)	Diff T1-C	Diff T2-C
		Chi-squared test		**0**	**0**
		Median percentage bias		**3.301**	**3.587**

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Standard errors are not clustered at the school-level.

Impact on pupil attainment

We now discuss the impact of the two sets of incentive schemes on our primary outcomes (performance in GCSE Maths, English and Science). The next section discusses the impact on pupil effort tasks (a secondary outcome).

Before detailing our main results, it is important to recall what the design of the experiment and the treatment/control conditions imply for how we should interpret the estimated treatment effects. As summarised in Table 2, the key difference between treatment and control groups is the provision of incentives for pupil effort across four domains and bespoke feedback letters documenting pupil effort levels. This means that **we should interpret the treatment effects as the combined effects of incentives and feedback**. There are other differences, however, which could influence the interpretation of the estimated treatment effect. It is likely that some control schools did not monitor pupil effort levels which means that the estimated treatment effect also includes this to some extent. Schools may also have changed their existing provision of incentive schemes after their allocation to treatment or control group.

Table 9 shows the differences in our primary outcomes (GCSE Maths, English and Science) across each of the treatment and control groups, as well as the differences in our secondary attainment outcomes (overall measures of GCSE performance across all subjects). Threshold measures show the proportion of pupils attaining each benchmark, while points score measures are standardised across pupils at the national level. The differences between each of the treatment groups and the control group are all small and not statistically significant. There is some evidence to suggest a small positive impact of both sets of incentives on Maths and a small negative impact on English, but these estimates are not statistically significant. These results may be affected by the imbalances in overall measures of school quality in the previous year, however.

Table 9: Comparison of GCSE outcomes at pupil level

Characteristic	Control group (C)	Financial incentives (T1)	Group incentives (T2)	Diff T1-C	Diff T2-C
Threshold measures					
Achieved grade C+ in Maths	0.639	0.663	0.651	0.024	0.013
Achieved grade C+ in English	0.630	0.610	0.592	-0.020	-0.038
Achieved EBacc Science component	0.336	0.327	0.366	-0.008	0.030
5+ GCSE equivalents A*-C (with Eng and Maths)	0.536	0.541	0.52	0.004	-0.017

Characteristic	Control group (C)	Financial incentives (T1)	Group incentives (T2)	Diff T1-C	Diff T2-C
5+ GCSE only A*-C (with Eng and Maths)	0.406	0.396	0.423	-0.010	0.018
Points score measures					
GCSE Maths points	-0.300	-0.269	-0.227	0.031	0.073
GCSE English points	-0.251	-0.334	-0.312	-0.083	-0.061
GCSE Science points	-0.404	-0.440	-0.405	-0.035	0
Capped GCSE equivalent points – new system	-0.215	-0.340	-0.402	-0.125	-0.187

Note: Standard errors are clustered at the school-level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Point score measures are standardised at the national level at the pupil-level.

In Appendix Figure A1, we also compare the full distribution of average (capped) GCSE points scores across the treatment groups and the control group. There are few differences in the shape of each of the three distributions, though those for the treatment groups are shifted down slightly (that is, have slightly lower attainment at a given point in the distribution) from the control group.

We now turn to our estimates of each of the treatment conditions on pupil attainment. We focus on the results from our preferred specification (FILM), which estimates the treatment effect at the pupil level having controlled for pupil and school characteristics as well as Key Stage 2 results (the pre-test in this context). We also refer to Appendix tables, which show the estimated effect of the treatment effect when we use a range of different methodologies to control for pupil and school characteristics, across a range of different samples and at school level. We also now restrict analysis to cases where none of the outcomes of interest is missing,[11] which means that we lose 336 pupils from the analysis of the effects on pupil attainment (the raw differentials in the Appendix are thus very slightly different from those in Table 9).

Financial incentives treatment

Table 10 shows the estimated effect of the financial incentives treatment effect condition on our primary outcomes (GCSE Maths, English and Science) and a secondary attainment outcome (overall GCSE capped points scores). This is shown for our preferred methodology (FILM). We estimate that there are small positive effects of the financial incentives treatment across GCSE Maths and English, although neither is statistically significant. The estimate for Science is slightly negative, but statistically insignificant as well. The estimated effect for overall points scored is very close to zero, suggesting that any impact on effort in the incentivised subjects did not significantly affect effort or attainment in other subjects. Not only are all these estimates statistically insignificant, but we can rule out medium-to-large positive effects (the upper end of the 95% confidence intervals are 0.12-0.16). Each of the estimated treatment effects is also consistent with negative effects. Therefore these estimates imply there is no evidence to suggest a positive effect of the financial incentives treatment on GCSE results: if there is an effect, it is unlikely to be large.

[11] This could be due to absence on the day of the test or pupils not entered for one of these three subjects.

Table 10: Estimated impact of financial incentives on primary GCSE outcomes

Outcome	Effect size	95% Confidence interval	N - Schools	N - Pupils
GCSE Maths points	0.037	(-0.06, 0.13)	48	7,730
GCSE English points	0.022	(-0.08, 0.12)	48	7,730
GCSE Science points	-0.058	(-0.27, 0.16)	48	7,730
Capped average point score	0.006	(-0.12, 0.13)	48	7,730

Note: Effect estimated using FILM controlling for pupil and school characteristics. Standard errors are clustered at the school level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 5% level (p<0.05). Point score measures are standardised at the national level at the pupil level.

In Appendix A, we report the results of various robustness checks to confirm these results. Here, we summarise the results and their implications. Appendix Table A4(a) shows that the estimates of the impact of the financial incentives on GCSE scores are very similar when we don't control for any pupil and school characteristics and when we do so using different methodologies (OLS or Kernel Matching). The similarity of the impact estimates across methodologies gives us greater confidence in the results and suggests the differences in pre-test characteristics observed in Tables 7 and 8 are not materially affecting our impact estimates.[12]

The matching estimators also allow us to undertake various diagnostic tests on how comparable the treatment and control groups are before and after matching. Figure A2(a) shows that although there are noticeable differences across some pupil and school characteristics across groups, these are much reduced post-matching. Table A4(a) also reports the results of a likelihood ratio test, with the null hypothesis that the groups are still imbalanced after matching. The p-value of 0.00 suggests we would still reject such a hypothesis and that the groups are still imbalanced. However, the fact that the estimates do not change radically when we control for pupil and school characteristics makes us doubt the extent to which this is materially biasing our estimates of the treatment effect.

Appendix Table A5(a) shows the estimated effect of financial incentives across the same range of outcomes based on school-level analysis. The estimated effects are similar in absolute value, though are slightly more positive. Indeed, after controlling for pupil and school characteristics, we observe significant and positive effects on Maths GCSE points, whether pupils gained a grade C or above in Maths and whether pupils gained 5 or more GCSEs at A*-C (including English and Maths).

Appendix Table A6 undertakes some limited sub-group analysis by examining whether the effect of the treatments is different when focusing just on pupils eligible for FSM or pupils with low prior attainment. The top half of the table relates to financial incentives. Estimates for pupils eligible for FSM are similar to those for all pupils and not statistically significantly different from zero. However, the estimates for pupils with low levels of prior attainment are much larger for Maths and English, although not statistically significant.

In summary, our main estimates provide no clear evidence that financial incentives led to improvements in our primary outcomes (Maths, English and Science) and we can rule out medium-to-large effects.

[12] This is despite the fact that the mean standardised difference in the propensity score (over 1) is outside the range suggested by Rubin (2007) for linear adjustment to be valid. Rubin argues that the mean difference should be less than 0.25 and the ratio of the variances should be between 0.5 and 2.

Event incentives treatment

Table 11 shows the estimated impact from our preferred specification of the event incentives treatment on the same set of outcome: our primary outcomes (GCSE Maths, English and Science); and capped GCSE points scored (a secondary outcome). The estimated effect for Maths is positive and statistically significant (though only at the 10% level). Estimates for English and capped average point score are also positive, though not statistically significant. For Science, estimates are negative and not statistically significant. However, in all cases the upper figures in the 95% confidence intervals imply that we can rule out large effects (less than 0.13-0.17). The overall pattern is thus quite similar to that for financial incentives. However, the estimates for Maths are clearly larger and are statistically significant at the 10% level. This provides some suggestion that the event incentive treatment positively affected Maths attainment, equivalent to around one sixth of a GCSE grade.

Table 11: Estimated impact of event incentives on primary GCSE outcomes

Outcome	Effect size	95% Confidence interval	N - Schools	N - Pupils
GCSE Maths points	0.084	(-0.01, 0.17)	48	7,980
GCSE English points	0.042	(-0.08, 0.16)	48	7,980
GCSE Science points	-0.063	(-0.25, 0.13)	48	7,980
Capped average point score	0.054	(-0.06, 0.17)	48	7,980

Note: Effect estimated using FILM controlling for pupil and school characteristics. Standard errors are clustered at the school level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Point score measures are standardised at the national level at the pupil level.

We also performed the same robustness checks as for financial incentives. Appendix Table A4(b) shows that the estimates of the impact of the event incentive treatment on GCSE scores change only very marginally (in the majority of cases) after accounting for pupil and school characteristics in each specification, suggesting that the small differences observed across groups do not seem to have a material effect on our ability to produce reliable impact estimates. Importantly, the positive effects on Maths remain in all cases after controlling for pupil and school characteristics. There are also positive effects on the proportion gaining a GCSE grade C or higher in Maths, though again these are not statistically significant.

The effects are also very similar if we estimate the treatment effects at school level (Table A5(b)), though none of estimated effects is statistically significant.

Unfortunately, diagnostic checks create some grounds for concern as to whether the event treatment and control groups are well balanced. Even after matching, Figure A2(b) shows that there are still sizeable differences across groups along a number of pre-test characteristics[13] (particularly the school-level characteristics such as whether schools are in London and prior school quality). It is reassuring to see that the estimated effects based on matching are similar to those for other methods. However, the size of the imbalance and the fact that matching cannot fully adjust for this imbalance makes one question the extent to which the event incentive treatment and control group are truly comparable.

We analysed whether the effects of the event incentive treatment varied across sub-groups based on our preferred FILM specification (Appendix Table A6). This shows that the estimated effects for pupils eligible for FSM are similar to the overall estimates and not statistically significantly different from zero.

[13] In seeking to re-weight the control group to look more like the treatment group, matching improves some of these differences, but still leaves some noticeable differences and makes some differences worse. This is a fundamental problem relating to the absolute difference in the size of the differences across some school characteristics and the small number of schools involved.

However, the estimated effects on Maths are larger and statistically significant for pupils with low levels of prior attainment. The estimated effect size of 0.138 is equivalent to about one quarter of a GCSE grade.

In summary, there is some evidence to suggest that the event incentives have had a positive effect on Maths GCSE scores, with a larger positive effect on those with low levels of prior attainment. Estimated effects are small and generally not statistically significant for other subjects,

Robustness checks

We also perform a number of robustness checks on our preferred sets of estimates in Tables 10 and 11. First, we examine whether the estimated effects differ if we exclude the 6 remainder schools who could not be placed into triplets (column (1)). The estimates are slightly more positive in this case, but the overall patterns are almost identical. The second and third column further show that our estimates are largely unchanged if we restrict analysis to the 15 randomised triplets and when we include dummy variables for the triplets, respectively.

The last column of Appendix Table A7 shows that our estimates of the treatment effects become somewhat more positive if we exclude control schools who failed to submit effort data. This suggests that control schools who did submit effort data actually had *lower* outcomes than those who failed to submit effort data (after controlling for pupil and school characteristics). This could be the case if monitoring had negative impacts on outcomes, but could also have resulted from schools who failed to submit data being different in unobservable ways.

Estimated impact on effort tasks

Financial and non-financial incentives that reward pupil effort are hypothesised to affect pupil attainment if two conditions are met: first, that pupil effort is increased as a result of the provision of incentives; second, that increased pupil effort has a positive impact on pupil attainment. This section explores the first condition: that pupil effort is increased as a result of the provision of each incentive scheme.

It is important to note the slight differences in the structure of the financial and event incentives, which affect the interpretation of the evidence presented in this section, which were discussed previously and summarised in Table 2. First, the weight applied to rewarding effort in classwork and homework was three times that for attendance and behaviour. We may therefore expect to see a larger impact of each incentive scheme on classwork and homework. Second, the event incentive scheme had a discrete threshold below which there was no reward for effort, while it was still possible to gain from effort in the financial incentive scheme below the equivalent threshold. We may therefore expect to see a larger impact of the financial incentive scheme across all measures.

The sample of pupils and schools varies slightly from the analysis of attainment outcomes in the previous section. This is because 18 of the 33 control schools did not submit behaviour data. The sub-sample of control schools that submitted behaviour data is likely to be a representative sample of the original control group: robustness checks presented in Appendix Table A7 show that the impact of the financial and event incentive treatment on pupil attainment is largely similar when either set of control schools is used. Also, the characteristics of these schools presented in Appendix Table A3 show few significant differences between each set of control schools, although there could be unobservable differences between the two groups.

A final note is that pupils in control schools were not provided with feedback letters, although equivalent monitoring took place in the subset of control schools that submitted behaviour data. As summarised in Table 2 and discussed in the surrounding text, it is therefore not possible to distinguish the impact of providing feedback letters about pupil effort from the provision of incentives.

The impact of financial incentive treatment and event incentive treatment on pupil effort

Figure 3 shows the estimated impact of financial incentives on pupil attendance. The darkest line represents the proportion of pupils in the control group that met the overall attendance target, in each half-term. Half-terms are represented by the numbers from 1 to 4 along the x-axis, which correspond to September-October, October-December, January-February and February-April half-terms. The lightest line represents the proportion of pupils in schools that were part of the financial incentive treatment group that met the overall target, and the slightly darker line represents the proportion of pupils in schools that were part of the event incentive treatment. The lines represent the proportion of pupils in each group that met the target, conditional on pupil characteristics (derived from our preferred FILM specification).

Figure 3 shows that the proportion of pupils achieving the overall attendance threshold declines slightly across the school year for each group of schools. The proportion of pupils achieving this threshold is generally slightly higher in schools with financial incentive treatment, while pupils in schools with an event incentive treatment have broadly the same effort level across all half-terms as pupils in control schools. Differences are not statistically significant, however, and therefore there is no strong suggestion that the incentive schemes implemented improve pupils' attendance at school. This suggests that any gains in attainment would be more likely to occur through effort while at school, rather than increased time in school.

Figure 4 shows a similar picture for pupil behaviour; pupils in schools with either incentive scheme are not significantly more likely to meet the behaviour target threshold than pupils in control group schools (although the proportion meeting the behaviour target in the event incentive treatment group is higher than in the control group). Attendance and behaviour thresholds receive a lower weight in each incentive scheme than classwork and homework. We might therefore expect to see a larger response of pupil effort in these domains.

Figure 5 and Figure 6 present the proportion of pupils in each treatment and control group that meet the threshold for classwork and homework respectively. Pupils in the financial incentive treatment group are significantly more likely to meet the threshold for classwork in the second half-term (before the Christmas break). A higher proportion of pupils in the event incentive treatment group meet the classwork target in this term than in the control group, but the difference is not statistically significant. Providing that teachers' assessment of classwork remains consistent across half-terms between treatment and control schools, this provides some evidence that incentive schemes can improve pupils' effort in classwork. The comparison between treatment and control groups in Figure 3 suggests that effort in classwork increases slightly for pupils in treatment groups and declines slightly for pupils in control groups in this half-term. This suggests that the incentive schemes are either particularly salient in this half-term (before the Christmas break) and/or that teachers in schools with incentive schemes retain more influence over the level of classwork of their pupils. Appendix Table A11 shows that the significant differences for the financial incentive treatment group are present for each subject as well as in the overall threshold measure for classwork.

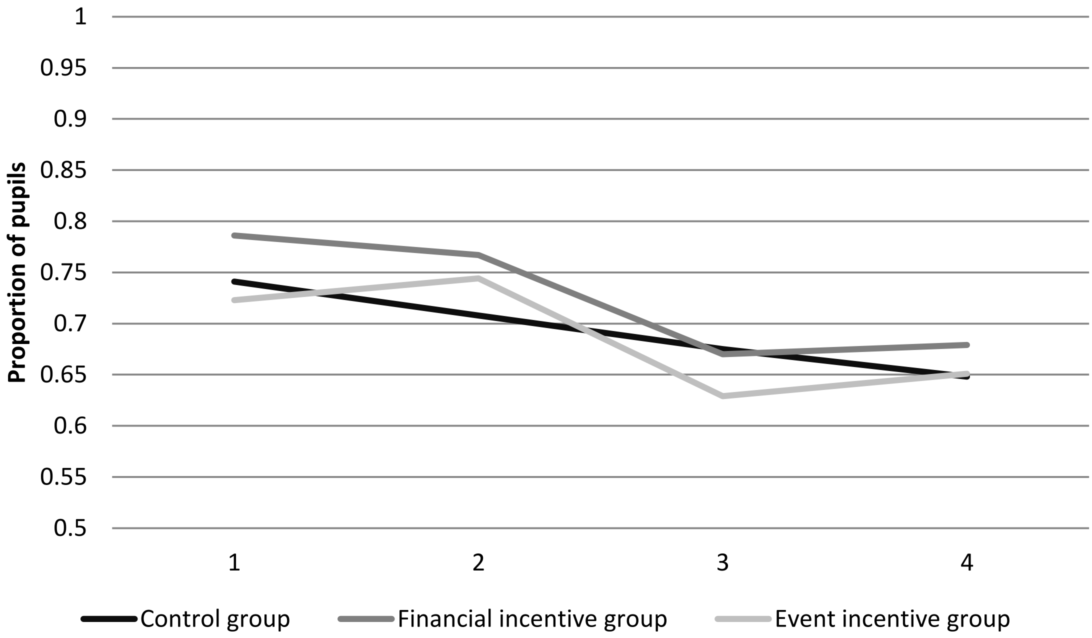
Figure 6 suggests that the salience of incentive schemes around the Christmas break does not translate into effort in homework, where pupils in both treatment groups are not significantly more likely to meet the threshold than the control group. As both homework and classwork have a larger weight in determining each pupil's eventual reward than attendance and behaviour in both incentive schemes, and there is no significant difference in homework, it is unlikely that the higher performance of pupils in classwork is due to this mechanism.

Appendix Tables A8 to A11 show the impact of each treatment on the effort of two sub-groups of pupils: pupils eligible and registered for free school meals, and pupils with low prior attainment (defined as achieving below the expected level of attainment in English, Maths or Science at the end of primary school).[14] These results find that the impact of incentives is broadly similar for these sub-groups of pupils: the response to incentives appears slightly more positive for pupils eligible and registered for free school meals (especially in behaviour and classwork), although the differences between this group of pupils and the main sample are not statistically significant.

[14] Unfortunately it was not possible to estimate these results for the sub-set of schools which were originally assigned to a triplet by the project team. This was because only 5 control schools that were originally assigned to a triplet submitted data on pupil effort to the project team. The number of schools with both information about pupil effort and original triplet assignment is too small to form the basis for a reasonable analysis.

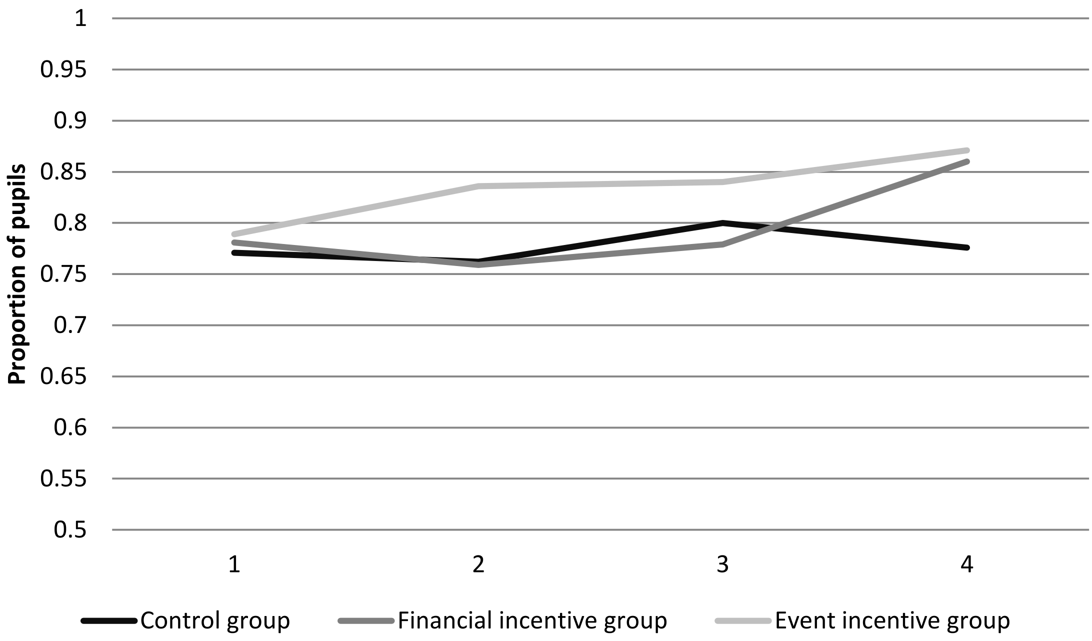
This section provides evidence that three of the four domains of pupil effort that were incentivised were not significantly affected by either the financial or event incentive schemes. Pupil classwork remained at a similarly high level across the school terms in both treatment groups, however, while classwork in the control group declined slightly in the half-terms around the Christmas break. This provides some evidence that the incentives for pupil effort, perhaps in addition to the salience of the rewards around this time of year and the positive influence of teachers in the classroom, can improve (or maintain) levels of pupil effort in classwork. This finding suggests that any positive impact of either incentive scheme on pupil attainment is likely to be through the impact of increased (or maintained) pupil effort in classwork, which is present across all subjects that were incentivised. Slight differences in the impact of the treatment groups on eventual pupil attainment across these three subjects may therefore reflect differences in the relationship between effort in classwork and attainment in Maths, English and Science (as found by Levitt et al. (2011) and Bettinger (2010)).

Figure 3: The proportion of pupils achieving the overall attendance threshold



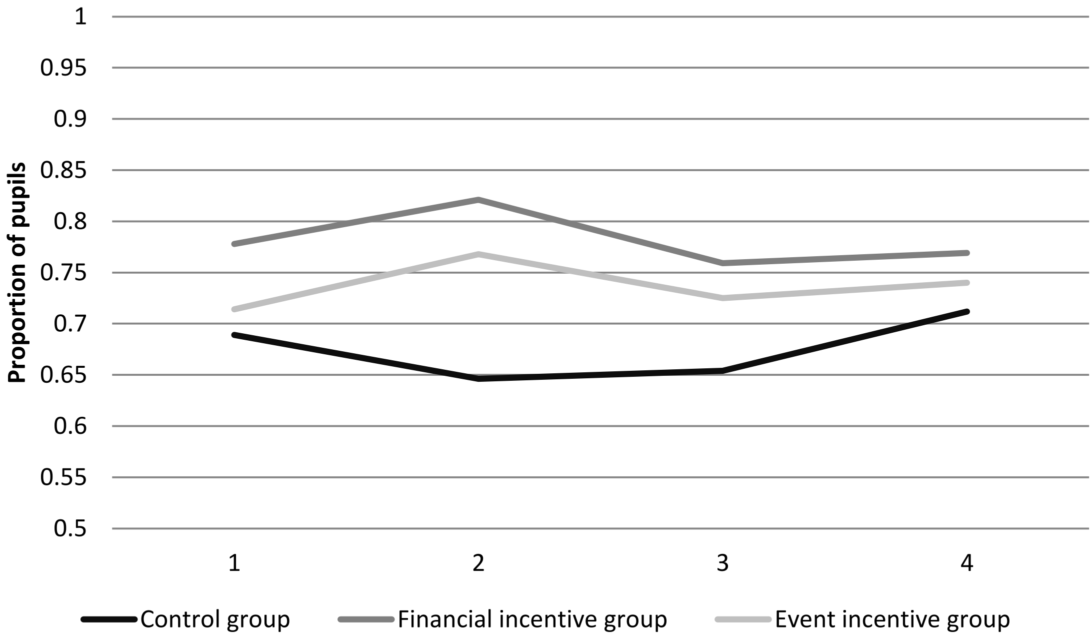
Note: differences between treatment and control groups are not statistically significant. Conditional proportions are calculated from two separate probit regressions where the specification matches the FILM specification reported in our main results (Table 9 and Table 10).

Figure 4: The proportion of pupils achieving the overall behaviour threshold



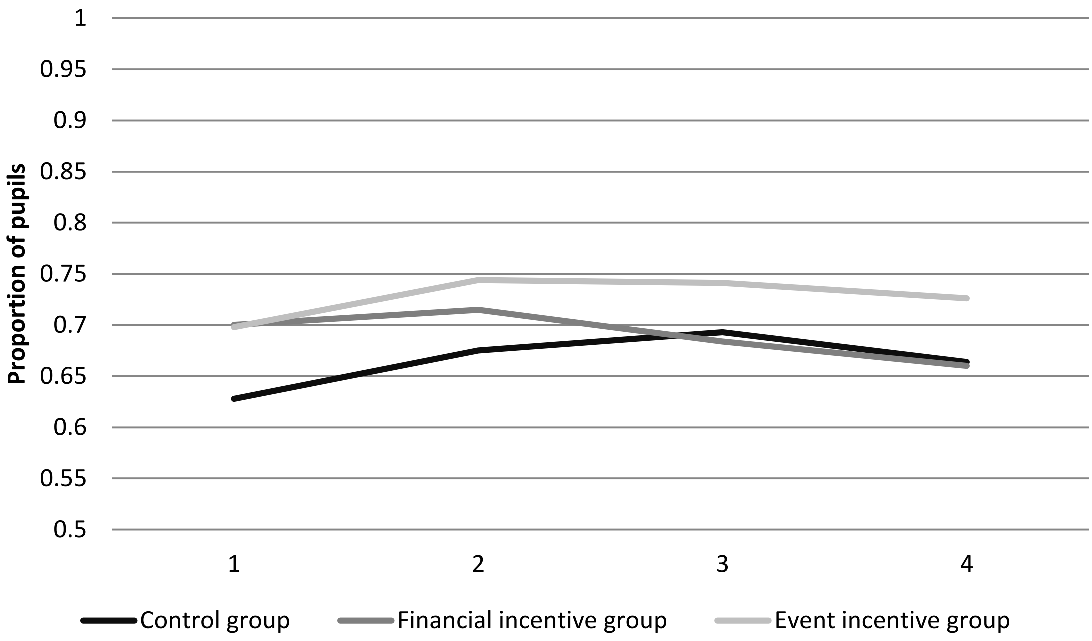
Note: differences between treatment and control groups are not statistically significant. Conditional proportions are calculated from two separate probit regressions where the specification matches the FILM specification reported in our main results (Table 9 and Table 10).

Figure 5: The proportion of pupils achieving the overall classwork threshold



Note: differences between treatment and control groups are statistically significant in the 2nd and 3rd term. Conditional proportions are calculated from two separate probit regressions where the specification matches the FILM specification reported in our main results (Table 9 and Table 10).

Figure 6: The proportion of pupils achieving the overall homework threshold



Note: differences between treatment and control groups are not statistically significant. Conditional proportions are calculated from two separate probit regressions where the specification matches the FILM specification reported in our main results (Table 9 and Table 10).

Cost

Given that the results were generally statistically insignificant, it is not appropriate to do a full cost-benefit analysis. However, we can detail the costs of implementing the schemes in this intervention.

The cost of the financial incentives represented a potential outlay of £320 per pupil if they met all their targets, Given the number of pupils who did meet their targets, the average cost per pupil was around £225 per pupil.

The cost of the event incentive treatment represented the budget given to schools for organising an event at the end of each term. The budget allocated by the project team was about £80 per pupil to cover both terms.

All schools were offered £2,000 to cover the expected costs of monitoring, though it is uncertain whether this is an accurate reflection of the true cost.

One important caveat to these figures is that there is no inherent cost to 'incentives' as it will depend on the level of incentives chosen in each case. Schools could, for instance, choose lower level incentives and significantly reduce the expected cost. This might also be an interesting avenue for future research. Lower level incentives across more schools might produce more precise estimates.

Process evaluation

There was no process evaluation commissioned as part of this independent evaluation. Instead, the project team conducted their own semi-structured interviews with staff members and pupils at the schools. The main findings from these semi-structured interviews are briefly summarised below.

- **Monitoring pupil effort and collecting data –** The project team reported that schools were initially concerned about the burden of data collection necessary for this experiment. However, the system developed by the project team was relatively simple: the project team provided significant assistance and a £2,000 payment was also made to cover the extra cost. In total, 45 of the 63 schools submitted effort data as part of the experiment. All of the 18 schools who dropped out were in the control group, suggesting that the perceived costs of collecting the data did lead to some schools dropping out when they perceived no benefit. Therefore, schools considering implementing such a scheme should make sure they are well prepared to collect the necessary effort data.

- **Paying financial incentives –** Originally, the project team aimed to pay the financial incentives into escrow bank accounts and deduct money as pupils missed targets (directly incorporating loss aversion). However, no bank was willing to participate in the experiment on this basis. This meant that the financial incentives had to be paid via cheques from the University of Bristol. If this system of incentives were taken up by individual schools, a new method would be needed to pay pupils their rewards.

- **Organising events –** In the non-financial incentives group, schools reported some difficulties in organising the end of term events. Deposits for travel had to be paid up-front, even though schools were uncertain how many pupils would meet their targets. Assistance was provided by the project team to ease this concern, both financially and organisationally. Furthermore, in this experiment about 60% of pupils were eligible to attend the event in each term. This figure should help schools (in a similar context to the schools in the intervention) planning events if they were to implement a similar scheme

- **Reported impact –** The project team asked schools whether they were planning to continue with this scheme in the future. Despite the lack of funding, they report that 5 treatment schools were planning to continue with the incentives.

Conclusion

Limitations

The main limitations of this evaluation are a potential lack of balance between treatment and control groups, the combination of incentives with feedback (which means we cannot estimate their individual effects in isolation), and a general lack of power as a result of the relatively small sample size.

There is **some evidence of an imbalance of pupil and school characteristics** across control and the two treatment groups, particularly the event treatment. While this doesn't seem to influence the findings substantively, there is some question over whether the groups can reasonably be compared.

A further limitation is the **combination of different incentives with other differences across treatment and control groups**. Table 2 in the methodology section summarises these differences. The main one is that pupils in treatment groups were also provided with feedback letters, as well as incentives for effort. We can therefore only estimate the combined effect of incentives and feedback. There could be other differences as well: the extent of monitoring is likely to have been lower in control schools, as about 18 of the 33 control schools dropped out from the intervention (by not monitoring and reporting pupil's effort). This could bias our estimates of the impact of the treatment on effort tasks (if the schools that stopped providing effort data are somehow different). We partially address such concerns, however, by showing that pupil and school characteristics do not vary significantly across control schools that dropped out and complied with the intervention. The estimated impact of each incentive treatment is not lower when we drop controls who did not submit effort, though it is higher, suggesting that schools who dropped out may be different in unobservable ways.

The final limitation of the current intervention was its size. The minimum detectable effect size was 0.219, which is clearly larger than the effect sizes found in the only comparable study (Fryer (2011) finds effect sizes of 0.01 for Reading and 0.08 for Maths). Having said that, our actual estimates of the effects of the treatment are relatively small and we are able to rule out medium-to-large positive effects. Therefore, any future experiments probably need to have slightly more power (i.e. higher numbers of schools), but may also be best focused on lower costs variants (e.g. focusing on non-financial incentives or smaller scale financial incentives).

Interpretation

There is some evidence to suggest that non-financial (event) incentives improved Maths scores at GCSE, though the effect is likely to be small – equivalent to about one sixth of a GCSE grade in Maths, and this result is significant only at the 10% level. There is also a positive effect of financial incentives, though smaller and not statistically significant. However, for both sets of incentives, there is a more positive effect of incentives on Maths scores for pupils with low levels of prior attainment – equivalent to about one quarter of a grade in GCSE Maths, although this is only significant at the 10% level for the financial incentive treatment. Across English and Science, there is no evidence of an effect of incentives on pupil attainment and we can again rule out medium-to-large effects.

There is some evidence that the provision of financial incentives contributed to higher pupil effort in classwork in the half-terms surrounding the Christmas break. This is true in all subjects, which suggests that the finding is not a spurious correlation. There is a slightly smaller positive effect of non-financial incentives. We tentatively conclude from this that the salience of the incentive scheme in the classroom is higher (especially at particular times of the year), and may suggest that teachers have otherwise relatively few mechanisms with which to incentivise pupil effort in the classroom. There is no statistically significant impact of either incentive scheme on pupil effort in homework, however, which had the same importance in the design of the incentive scheme, or on behaviour and attendance at school.

Combined with the results for attainment, this suggests that higher pupil effort in classwork possibly translates into higher pupil attainment in Maths. However, there is no evidence that this link between effort and attainment exists for Science or English. This could be the case if there is a different relationship between classwork and test scores across subjects (e.g. if attainment in Maths depends more on effort in classwork than other subjects).

Future research and publications

A future intervention could test the effectiveness of different levels of incentives. For instance, one could test lower level (and therefore cheaper) financial incentives with a greater number of schools, which could improve the precision of estimates. One could also consider increasing the generosity of the event incentives, given that this scheme was cheaper than the financial incentive scheme and appeared to have similar positive effects on Maths scores.

Given that the estimated impact on pupil attainment is larger for Maths and for pupils with low level of prior attainment, further work might be required to determine whether there are any adverse effects to rewarding a single subject (e.g. Maths) or certain groups of pupils (e.g. those with low levels of prior attainment).

Previous literature demonstrates some positive impact of providing incentives to pupils, for either indirect inputs such as effort or direct outputs such as test scores. It is unclear whether these short-term impacts affect long-term outcomes, however, such as progression to further education or employment. Future research should therefore include the scope for assessing the long-term implications of providing incentives to pupils, which could be positive if increased attainment leads to higher outcomes or negative if the provision of incentives permanently displaced some intrinsic motivation for effort. We are aware that EEF will be tracking pupils in this project through the EEF data archive, which will allow researchers to look at performance beyond Key Stage 4 and potentially into higher education.

Previous research has shown that a decline in pupil effort around the time of high-stakes exams negatively affects a pupil's performance (Burgess et al, 2011). This suggests that incentive schemes that relate to effort around the period of assessment may be both more salient and lead to greater improvements in pupil attainment. As noted above, however, it is crucial to assess whether any short-term improvements in exam performance translate into improved long-term outcomes. One could also consider whether incentives are more effective earlier on in pupils' school careers. It may well be that additional effort in Year 11 is simply too late to make a material effect on pupil attainment.

References

Angrist, J. and V. Lavy (2009) 'The effects of high stakes high school achievement awards: Evidence from a randomized trial', *American Economic Review*, 99(4), 1384–1414.

Ashraf, N., Bandiera, O. and S.S. Lee (2014) 'Awards unbundled: Evidence from a natural field experiment,' *Journal of Economic Behavior & Organization*, Elsevier, vol. 100(C), 44–63.

Baumert, J. and A. Demmrich (2001) 'Test motivation in the assessment of student skills: The effects of incentives on motivation and performance', *European Journal of Psychology of Education*, 16(3), 441–462.

Besley, T. and M. Ghatak (2008) 'Status incentives', *American Economic Review*, 98(2), 206–11.

Bettinger, E.P. (2012) 'Paying to learn: The effect of financial incentives on elementary school test scores', *The Review of Economics and Statistics*, 94(3), 686–698.

Braun, H., Kirsch, I. and K. Yamamoto (2011) 'An experimental study of the effects of monetary incentives on performance on the 12th-Grade NAEP Reading Assessment', Teachers College Record.

Bruhn, M., and D. McKenzie (2009) 'In pursuit of balance: Randomization in practice in development field experiments', *American Economic Journal: Applied Economics*, 1(4), 200–232.

Burgess, S. and M. Ratto (2003) 'The role of incentives in the public sector: Issues and evidence', *Oxford Review of Economic Policy*, 19, 285–300.

Deci, E., Koestner, R. and R. Ryan (1999) 'A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation', *Psychological Bulletin*, 125, 692–700.

Fryer, R. (2011) 'Financial incentives and student achievement: Evidence from randomized trials', *Quarterly Journal of Economics*, 126(4), 1755–1798.

Imbens, G.W. and J.M. Wooldridge (2009) 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature*, 47(1), 5–86.

Jackson, C. K. (2010) 'A little now for a lot later: A look at a Texas advanced placement incentive program', *Journal of Human Resources*, 45(3), 591–639.

Kosfeld, M. and S. Neckermann (2011) 'Getting more work for nothing? Symbolic awards and worker performance', *American Economic Journal: Microeconomics*, 3(3), 86–99.

Levitt, S., List, J., Neckermannm S. and S. Sadoff (2011) *The Impact of Short-Term Financial Incentives on Student Performance*, University of Chicago.

Metcalfe, R., Burgess, S. and S. Proud (2011) 'Using the England football team to identify the education production function: Student effort, educational attainment and the World Cup', CMPO Working Paper.

Muralidharan, K. and V. Sundararaman (2011) 'Teacher performance pay: Experimental evidence from India', *Journal of Political Economy*, 119, 39–77.

O'Neil, H.D. Jr., Sugrue, B., Baker, E.L. and S. Golan (1997) 'Final report of experimental studies on motivation and NAEP test performance', (CSE Tech. Rep. No.427), Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Rubin, D.B. (2007) 'The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials', *Statistics in Medicine*, 26(1), 20–36.

Tirole, J. and R. Bénabou (2006) ‘Incentives and prosocial behavior’, *American Economic Review*, 96(5), 1652–1678.

Wilson, D., Burgess, S. and A. Briggs (2011) ‘The dynamics of school attainment of England’s ethnic minorities’, *Journal of Population Economics*, 24(2), 681–700.

Appendix A – Tables and figures

Figure A1: Comparison of distribution of capped GCSE points scores across treatments and control group

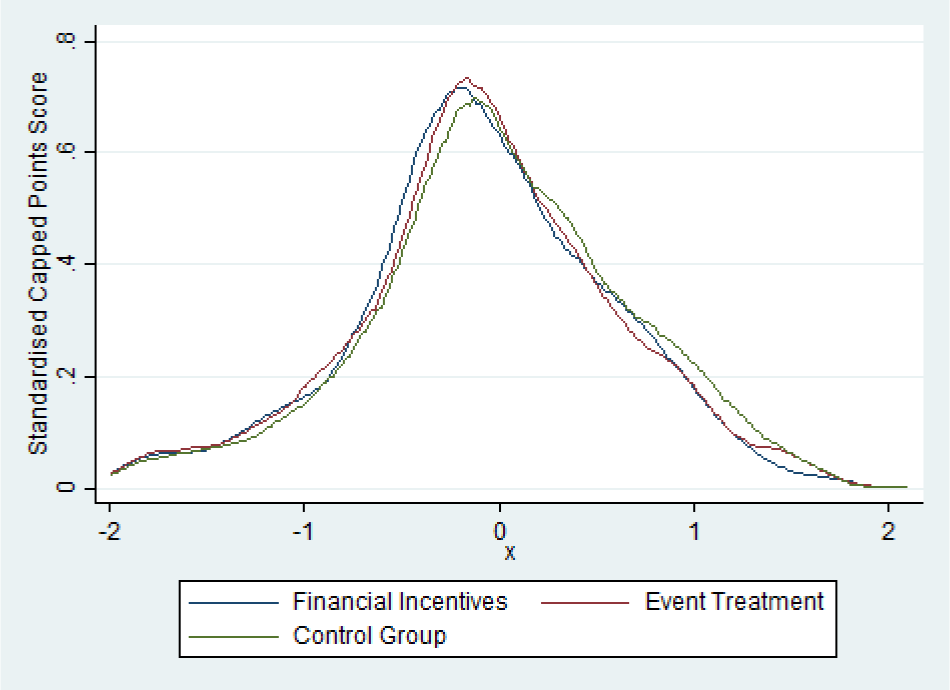


Figure A2(a): Differences in pupil and school characteristics between financial incentives and control group, pre- and post-matching

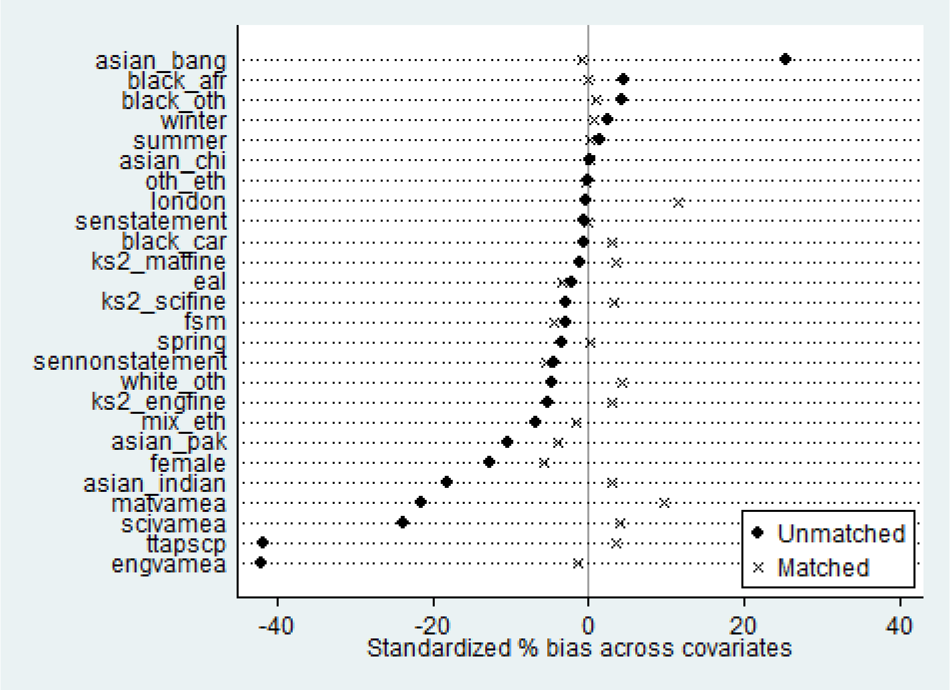


Figure A2(b): Differences in pupil and school characteristics between event incentive treatment and control group, pre- and post-matching

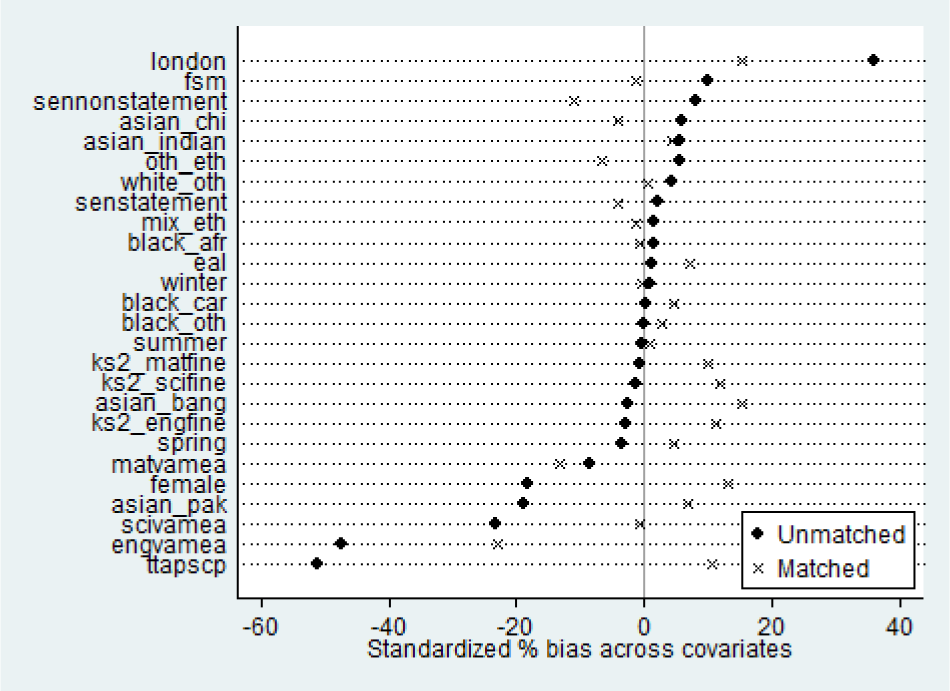


Table A1: Intra-class correlation across primary outcomes

	Without controls	With controls
GCSE Maths points	0.09	0.03
GCSE English points	0.13	0.04
GCSE Science points	0.20	0.14
Capped GCSE equivalents points - new system	0.11	0.04

Table A2: Outcomes of control schools – stayers vs dropouts

Characteristic	Stayers	Dropped out	Diff
Threshold measures			
Achieved Grade C+ in Maths	0.626	0.610	-0.016
Achieved Grade C+ in English	0.616	0.598	-0.018
Achieved EBacc Science component	0.311	0.298	-0.013
5+ GCSE equivalents A*-C (with Eng and Maths)	0.524	0.500	-0.024
5+ GCSE only A*-C (with Eng and Maths)	0.396	0.354	-0.042
Points score measures			
GCSE Maths points	-0.342	-0.385	-0.043
GCSE English points	-0.303	-0.326	-0.023
GCSE Science points	-0.412	-0.510	-0.097
Capped GCSE equivalents points - new system	-0.236	-0.298	-0.062

Note: Standard errors in brackets and are clustered at the school level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01).

Table A3: Characteristics of control schools – stayers vs dropouts

Characteristic	Stayers	Dropped out	Diff
Student demographics			
Percentage of year group who are female	0.563	0.540	-0.023
Percentage of year group who are non-white	0.670	0.605	-0.065
Percentage of year group who are of black ethnicity	0.186	0.229	0.042
Percentage of year group who are of Asian ethnicity	0.281	0.113	-0.168*
Percentage of year group eligible for FSM	0.429	0.359	-0.070
Percentage of year group with SEN statement	0.017	0.026	0.009
Percentage of year group with EAL	0.524	0.378	-0.146
Structural characteristics			
London	0.333	0.500	0.167
Single sex school	0.200	0.111	-0.089
Academy	0.267	0.333	0.067
Has a sixth form	0.533	0.722	0.189
School size	901.400	926.556	25.156
Cohort size	168.133	168.389	0.256
Has existing incentive scheme	0.400	0.222	-0.178
Prior performance			
English Baccalaureate Maths Value Added measure	1001.000	999.911	-1.089
English Baccalaureate English Value Added measure	1000.913	1000.622	-0.291
English Baccalaureate Science Value Added measure	1000.280	999.489	-0.791
Capped points score at GCSE and equivalent	328.767	331.039	2.272
Chi-squared test			0.035
Median percentage bias			24.799

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01).

Table A4(a): Impact analysis at pupil level – financial incentives vs control group

Characteristic	Raw comparison	OLS/Probit	FILM	Kernel matching
Threshold achievement measures				
Achieved Grade C+ in Maths	0.027	0.044*	0.026	0.044
	[0.036]	[0.021]	[0.021]	[0.039]
Achieved Grade C+ in English	-0.02	0.028	0.024	0.039
	[0.050]	[0.026]	[0.023]	[0.047]
Achieved EBacc Science component	-0.009	0.016	-0.003	0.018
	[0.054]	[0.030]	[0.032]	[0.042]
5+ GCSE or equivalents A*-C (with E&M)	0.006	0.042	0.032	0.053
	[0.046]	[0.023]	[0.020]	[0.044]
5+ GCSE only A*-C (with E&M)	-0.009	0.019	0.007	0.026
	[0.058]	[0.023]	[0.021]	[0.041]
Points score measures (standardised)				
GCSE Maths points	0.038	0.056	0.037	0.082
	[0.106]	[0.052]	[0.047]	[0.094]
GCSE English points	-0.095	0.026	0.022	0.040
	[0.118]	[0.058]	[0.051]	[0.108]
GCSE Science points	-0.037	-0.022	-0.058	-0.016
	[0.140]	[0.111]	[0.109]	[0.138]
Capped GCSE equivalents points	-0.107	0.019	0.006	0.010
	[0.089]	[0.066]	[0.064]	[0.113]
Matching diagnostics			**Pre**	**Post**
Median absolute standardised bias			4.374	2.990
Chi-squared test (p-value)			0	0
Absolute difference in standardised propensity score			1.015	0.020
Ratio of variance of propensity score			1.074	1.002

Note: Standard errors in brackets and are clustered at the school level. Matching standard errors are calculated using the bootstrap method and are clustered at the school level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Continuous points score measured are standardised at the national level at the pupil level.

Table A4(b): Impact analysis at pupil level – event incentive treatment vs control group

Characteristic	Raw comparison	OLS/Probit	FILM	Kernel matching
Threshold achievement measures				
Achieved Grade C+ in Maths	0.017	0.033	0.036	0.026
	[0.032]	[0.022]	[0.022]	[0.056]
Achieved Grade C+ in English	-0.031	0.021	0.016	0.004
	[0.044]	[0.026]	[0.028]	[0.061]
Achieved EBacc Science component	0.033	0.062*	0.062*	0.078
	[0.053]	[0.027]	[0.030]	[0.066]
5+ GCSE or equivalents A*-C (with E&M)	-0.011	0.026	0.026	0.010
	[0.040]	[0.022]	[0.022]	[0.058]
5+ GCSE only A*-C (with E&M)	0.023	0.047*	0.051*	0.069
	[0.053]	[0.022]	[0.022]	[0.056]
Points score measures (standardised)				
GCSE Maths points	0.077	0.083	0.084	0.086
	[0.092]	[0.049]	[0.046]	[0.130]
GCSE English points	-0.069	0.048	0.042	0.039
	[0.117]	[0.059]	[0.060]	[0.149]
GCSE Science points	0.008	-0.036	-0.063	0.041
	[0.153]	[0.090]	[0.096]	[0.173]
Capped GCSE equivalents points	-0.140	0.046	0.054	0.090
	[0.092]	[0.057]	[0.060]	[0.163]
Matching diagnostics			**Pre**	**Post**
Median absolute standardised bias			3.939	5.619
Chi-squared test (p-value)			0	0
Absolute difference in standardised propensity score			1.171	0.023
Ratio of variance of propensity score			1.240	1.011

Note: Standard errors in brackets and are clustered at the school level. Matching standard errors are calculated using the bootstrap method (500 repetitions) and are clustered at the school level. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Continuous points score measured are standardised at the national level at the pupil-level.

Table A5(a): Impact analysis at school level – financial incentives vs control group

Characteristic	Raw comparison	OLS/Probit
Threshold achievement measures		
Achieved Grade C+ in Maths	0.025	0.054**
	[0.036]	[0.019]
Achieved Grade C+ in English	-0.022	0.023
	[0.048]	[0.035]
Achieved EBacc Science component	-0.011	0.025
	[0.055]	[0.033]
5+ GCSE or equivalents A*-C (with E&M)	0.004	0.047
	[0.044]	[0.026]
5+ GCSE only A*-C (with E&M)	-0.011	0.025
	[0.057]	[0.030]
Points score measures		
GCSE Maths points	0.035	0.089
	[0.104]	[0.050]
GCSE English points	-0.098	0.014
	[0.122]	[0.079]
GCSE Science points	-0.038	-0.01
	[0.153]	[0.118]
Capped GCSE equivalents points	-0.109	0.023
	[0.100]	[0.058]

Note: Standard errors in brackets. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Continuous outcomes are standardised at the national level at the pupil level.

Table A5(b): Impact analysis at school level – event incentive treatment vs control group

Characteristic	Raw comparison	OLS/Probit
Threshold achievement measures		
Achieved Grade C+ in Maths	0.018	0.015
	[0.034]	[0.019]
Achieved Grade C+ in English	-0.030	-0.013
	[0.047]	[0.035]
Achieved EBacc Science component	0.035	0.041
	[0.053]	[0.033]
5+ GCSE or equivalents A*-C (with E&M)	-0.010	-0.001
	[0.042]	[0.027]
5+ GCSE only A*-C (with E&M)	0.025	0.019
	[0.055]	[0.030]
Points score measures		
GCSE Maths points	0.080	0.038
	[0.101]	[0.051]
GCSE English points	-0.070	-0.027
	[0.118]	[0.080]
GCSE Science points	0.012	-0.078
	[0.147]	[0.120]
Capped GCSE equivalents points	-0.139	-0.008
	[0.097]	[0.059]

Note: Standard errors in brackets. * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Continuous outcomes are standardised at the national level at the pupil level.

Table A6: Differential impact of incentives across sub-groups (FILM specification)

	All pupils	Pupils eligible for FSM	Pupils with low attainment
Financial incentives			
GCSE Maths points	0.037	0.062	0.137
	[0.047]	[0.061]	[0.080]
GCSE English points	0.022	0.030	0.116
	[0.051]	[0.062]	[0.078]
GCSE Science points	-0.058	-0.014	-0.052
	[0.109]	[0.112]	[0.151]
Capped GCSE equivalents points	0.006	0.038	0.052
	[0.064]	[0.081]	[0.098]
Event incentives			
GCSE Maths points	0.084	0.091	0.138*
	[0.046]	[0.058]	[0.055]
GCSE English points	0.042	0.076	0.099
	[0.060]	[0.069]	[0.079]
GCSE Science points	-0.063	-0.094	-0.196
	[0.096]	[0.102]	[0.122]
Capped GCSE equivalents points	0.054	0.087	0.086
	[0.060]	[0.075]	[0.074]

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Standard errors are not clustered at the school level.

Table A7: Robustness of estimates across different samples

	All pupils	Dropping 'Type C' schools	Original triplets	With triplet dummy variables	With effort data
Financial incentives					
GCSE Maths points	0.037	0.059	0.042	-0.008	0.119*
	[0.047]	[0.056]	[0.047]	[0.039]	[0.055]
GCSE English points	0.022	0.077	0.063	-0.017	0.039
	[0.051]	[0.053]	[0.058]	[0.044]	[0.069]
GCSE Science points	-0.058	-0.106	-0.064	-0.113	0.079
	[0.109]	[0.117]	[0.098]	[0.075]	[0.082]
Capped GCSE equivalents points	0.006	0.022	0.032	-0.038	0.023
	[0.064]	[0.074]	[0.061]	[0.065]	[0.078]
Event incentives					
GCSE Maths points	0.084	0.106*	0.102	0.114**	0.164**
	[0.046]	[0.045]	[0.054]	[0.034]	[0.063]
GCSE English points	0.042	0.071	0.050	0.061	0.019
	[0.060]	[0.061]	[0.079]	[0.051]	[0.077]
GCSE Science points	-0.063	-0.082	-0.045	0.010	0.015
	0.054	0.077	0.103*	-0.034	0.089
Capped GCSE equivalents points	[0.060]	[0.057]	[0.057]	[0.037]	[0.071]

Note: * indicates that the difference in means is significant at the 5% level (p<0.05). ** at the 1% level (p<0.01). Standard errors are not clustered at the school level. The final specification focuses on the original set of 15 Type A triplets and includes dummy variables for each triplet.

Table A8: The impact of financial and effort treatments on pupil effort: attendance threshold

	Half-term 1	Half-term 2	Half-term 3	Half-term 4	Number of pupils
Financial incentive treatment					
All pupils	0.045	0.059	-0.005	0.031	4705
	[0.043]	[0.062]	[0.059]	[0.062]	
Pupils eligible and registered for free school meals	0.042	0.077	0.018	0.062	4705
	[0.048]	[0.070]	[0.067]	[0.067]	
Pupils with low prior attainment	0.030	0.060	-0.044	-0.025	4705
	[0.050]	[0.071]	[0.066]	[0.068]	
Event incentive treatment					
All pupils	-0.017	0.036	-0.046	0.003	4971
	[0.049]	[0.051]	[0.054]	[0.045]	
Pupils eligible and registered for free school meals	0.007	0.045	-0.032	-0.001	4971
	[0.054]	[0.058]	[0.055]	[0.053]	
Pupils with low prior attainment	-0.015	0.056	-0.058	-0.019	4971
	[0.053]	[0.067]	[0.062]	[0.060]	

Table A9: The impact of financial and effort treatments on pupil effort: behaviour threshold

	Half-term 1	Half-term 2	Half-term 3	Half-term 4	Number of pupils
Financial incentive treatment: overall					
All pupils	0.011	-0.003	-0.020	0.084	4705
	[0.047]	[0.053]	[0.047]	[0.056]	
Pupils eligible and registered for free school meals	0.046	0.041	0.020	0.106	4705
	[0.049]	[0.059]	[0.053]	[0.059]	
Pupils with low prior attainment	-0.012	-0.022	-0.050	0.063	4705
	[0.053]	[0.070]	[0.059]	[0.068]	
Event incentive treatment: overall					
All pupils	0.018	0.074	0.041	0.095	4971
	[0.052]	[0.061]	[0.065]	[0.072]	
Pupils eligible and registered for free school meals	0.043	0.115	0.113	0.126	4971
	[0.058]	[0.070]	[0.069]	[0.070]	
Pupils with low prior attainment	-0.024	0.069	0.033	0.054	4971
	[0.052]	[0.068]	[0.067]	[0.070]	
Financial incentive treatment: English (all pupils)	-0.018	-0.055	-0.049	-0.010	4705
	[0.031]	[0.032]	[0.031]	[0.033]	
Event incentive treatment: English (all pupils)	-0.004	0.026	-0.007	0.024	4971
	[0.035]	[0.033]	[0.037]	[0.032]	
Financial incentive treatment: Maths (all pupils)	0	-0.026	-0.031	0.041	4705
	[0.031]	[0.031]	[0.034]	[0.037]	
Event incentive treatment: Maths (all pupils)	-0.004	0.028	0.015	0.048	4971
	[0.039]	[0.045]	[0.047]	[0.057]	
Financial incentive treatment: Science (all pupils)	-0.027	-0.007	-0.012	0.041	4705
	[0.034]	[0.033]	[0.031]	[0.031]	
Event incentive treatment: Science (all pupils)	-0.009	0.013	-0.007	0.030	4971
	[0.038]	[0.043]	[0.051]	[0.043]	

Table A10: The impact of financial and effort treatments on pupil effort: classwork threshold

	Half-term 1	Half-term 2	Half-term 3	Half-term 4	Number of pupils
Financial incentive treatment: overall					
All pupils	0.089	0.176*	0.104	0.057	4705
	[0.067]	[0.069]	[0.077]	[0.075]	
Pupils eligible and registered for free school meals	0.134	0.208**	0.138	0.065	4705
	[0.075]	[0.080]	[0.080]	[0.080]	
Pupils with low prior attainment	0.072	0.174*	0.099	0.051	4705
	[0.076]	[0.088]	[0.089]	[0.090]	
Event incentive treatment: overall					
All pupils	0.025	0.122	0.070	0.028	4971
	[0.081]	[0.076]	[0.079]	[0.072]	
Pupils eligible and registered for free school meals	0.028	0.131	0.070	0.004	4971
	[0.083]	[0.078]	[0.076]	[0.068]	
Pupils with low prior attainment	-0.032	0.126	0.041	0	4971
	[0.087]	[0.080]	[0.081]	[0.079]	
Financial incentive treatment: English (all pupils)	0.079*	0.098	0.068	0.046	4705
	[0.038]	[0.054]	[0.049]	[0.042]	
Event incentive treatment: English (all pupils)	0.038	0.072	0.041	0.035	4971
	[0.044]	[0.057]	[0.051]	[0.043]	
Financial incentive treatment: Maths (all pupils)	0.060	0.133*	0.064	0.027	4705
	[0.047]	[0.054]	[0.051]	[0.045]	
Event incentive treatment: Maths (all pupils)	0.013	0.088	0.021	-0.004	4971
	[0.064]	[0.064]	[0.063]	[0.050]	
Financial incentive treatment: Science (all pupils)	0.050	0.120*	0.073	0.063	4705
	[0.044]	[0.049]	[0.042]	[0.049]	
Event incentive treatment: Science (all pupils)	-0.009	0.050	0.053	0.027	4971
	[0.049]	[0.048]	[0.045]	[0.046]	

Table A11: The impact of financial and effort treatments on pupil effort: homework threshold

	Half-term 1	Half-term 2	Half-term 3	Half-term 4	Number of pupils
Financial incentive treatment: overall					
All pupils	0.072	0.040	-0.010	-0.004	4705
	[0.070]	[0.070]	[0.075]	[0.083]	
Pupils eligible and registered for free school meals	0.104	0.065	0.008	0.032	4705
	[0.074]	[0.080]	[0.077]	[0.088]	
Pupils with low prior attainment	0.024	0.026	-0.039	-0.029	4705
	[0.083]	[0.080]	[0.088]	[0.091]	
Event incentive treatment: overall					
All pupils	0.070	0.069	0.047	0.062	4971
	[0.071]	[0.067]	[0.073]	[0.075]	
Pupils eligible and registered for free school meals	0.054	0.073	0.053	0.063	4971
	[0.075]	[0.073]	[0.073]	[0.073]	
Pupils with low prior attainment	0.023	0.083	0.016	0.023	4971
	[0.080]	[0.069]	[0.076]	[0.075]	
Financial incentive treatment: English (all pupils)	0.047	0.079	0.047	0.048	4705
	[0.043]	[0.049]	[0.056]	[0.058]	
Event incentive treatment: English (all pupils)	0	0.057	0.035	0.050	4971
	[0.046]	[0.054]	[0.051]	[0.052]	
Financial incentive treatment: Maths (all pupils)	0.026	-0.002	-0.038	-0.028	4705
	[0.053]	[0.050]	[0.052]	[0.052]	
Event incentive treatment: Maths (all pupils)	0.065	0.074	0.038	0.036	4971
	[0.055]	[0.051]	[0.056]	[0.054]	
Financial incentive treatment: Science (all pupils)	0.048	0.030	0.012	0.025	4705
	[0.052]	[0.051]	[0.046]	[0.050]	
Event incentive treatment: Science (all pupils)	0.028	-0.009	-0.016	0.027	4971
	[0.058]	[0.052]	[0.056]	[0.048]	

Appendix B – Definitions and measurement of effort

There were four domains of effort (attendance, behaviour, classwork, homework). These were defined and measured as follows across treatment and control groups:

Attendance – Attendance at school, not lessons, measured twice a day in registration by teachers (the official measure of attendance). The threshold for success was no unauthorised absences per half-term. This is a standard measure of attendance that is already recorded by schools.

Behaviour – An instance of poor behaviour was recorded on the basis of the student either (a) arriving late to a lesson (more than 5 minutes late); or (b) exhibiting behaviour in the lesson resulting in a sanction. Arriving late to a lesson could be judged objectively by the teacher and recorded by them. Exhibiting poor behaviour required some subjective judgement by the teacher and was recorded by the teacher in each lesson and centrally by the school. The threshold for success was no more than one instance of bad behaviour per subject per half-term (English, Maths and Science).

Classwork – Completion of work on time and at a level consistent with the student’s target (at a lower level only with strong acceptable justification). This was judged by the individual subject teacher across Maths, English and Science and recorded centrally by the schools. To be successful, work across all three subjects had to meet this standard each half-term.

Homework – Completion of work on time and at a level consistent with the student’s target (at a lower level only with strong acceptable justification). This includes participation in any out-of-school learning/homework club as required by the school. To be successful, work across all three subjects had to meet this standard each half-term.

Appendix C: Example Feedback Letter

NAME:	**«firstname» «secondname»**
PUPIL URN:	**«pupilupn»**
SCHOOL:	**«sch_name»**

Date: January 2013

This is your report card for the second half-term from October to December.

1. ATTENDANCE

Unauthorised absences per half-term:

TARGET: **0**
YOUR RESULT: You «text_a_thresholdattend» the target

2. BEHAVIOUR

Instances of poor behaviour per half-term:

TARGET: **No more than 1**
YOUR RESULT: You «text_a_thresholdbehave» the target

«text_b»

3. CLASSWORK

Class assignments completed on time and to a level consistent with your target grade:

TARGET: **ALL**
YOUR RESULT: You «text_a_thresholdclasswk» the target

4. HOMEWORK

Home assignments completed on time and to a level consistent with your target grade:

TARGET: **ALL**
YOUR RESULT: You «text_a_thresholdhomewk» the target

«text_c»

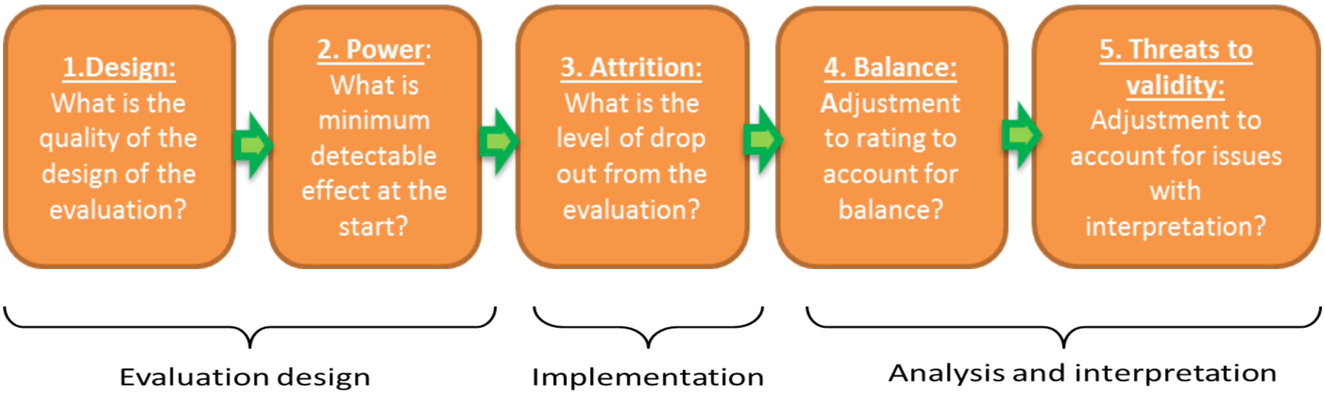
OVERALL OUTCOME for «firstname» «secondname»

Due to your «text_d»«amtlostorwon» of the £80 in your account. You will receive £«finreward» in a few days' time.

You still have £160 in your account to work for. Please «text_f» next half-term to achieve all your targets so you do not lose any of it.

Appendix D: EEF security rating summary

Security rating summary:



1.Design: What is the quality of the design of the evaluation?

2. Power: What is minimum detectable effect at the start?

3. Attrition: What is the level of drop out from the evaluation?

4. Balance: Adjustment to rating to account for balance?

5. Threats to validity: Adjustment to account for issues with interpretation?

Evaluation design

Implementation

Analysis and interpretation

Rating	1. Design	2. Power (MDES)	3. Attrition	4. Balance	5. Threats to validity
5 🔒	🟩 Fair and clear experimental design (RCT)	🟩 < 0.2	🟩 < 10%	Well-balanced on observables	No threats to validity
4 🔒	Fair and clear experimental design (RCT, RDD)	< 0.3	< 20%	¦	¦
3 🔒	Well-matched comparison (quasi-experiment)	< 0.4	< 30%	🟩	¦
2 🔒	Matched comparison (quasi-experiment)	< 0.5	< 40%	¦	🟩
1 🔒	Comparison group with poor or no matching	< 0.6	< 50%	↓	↓
0 🔒	No comparator	> 0.6	> 50%	Imbalanced on observables	Significant threats

The final security rating for this trial is 2 🔒. This means that findings are of modest security.

The trial was designed as a cluster randomized efficacy trial with the intention of recruiting 45 schools. 63 were achieved and the additional schools added to the control group resulting in a minimum detectable effect size of about 0.16 at randomization for attainment meaning the trial could still have achieved a maximum of 5 🔒. There was no attrition from the primary outcome of GCSE scores as this was obtained from the National Pupil Database. However, there was some significant imbalance at the baseline on pupil characteristics of FSM, ethnicity and prior attainment due to schools being imbalanced at randomization; this was controlled for in the analysis. The randomization was not done independently and there was a risk that some control schools implemented incentive schemes anyway.

Appendix E : Cost rating

Cost ratings are based on the approximate cost per pupil of implementing the intervention over one year. Cost ratings are awarded using the following criteria.

Cost	Description
£	*Very low:* less than £80 per pupil per year.
£ £	*Low:* up to about £170 per pupil per year.
£ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk